Face-to-face meeting of the LSST Science Advisory Committee Monday, August 13, 2018, Tucson, Arizona

SAC members present: Michael Strauss, Franz Bauer, Timo Anguita, Risa Wechsler, Niel Brandt, Josh Simon, Charles Liu, David Kirkby, Lucianne Walkowicz (remote), Jason Kalirai (remote), Anze Slosar (remote)

A number of LSST Project personnel were in attendance and took part in the discussion and made presentations, including Steve Kahn, Zeljko Ivezic, and others.

We discussed two main topics in this meeting:

-The draft data rights/data access document, prepared by Beth Willman, Melissa Graham, Wil O'Mullane and their colleagues. Melissa gave a presentation on this document (https://project.lsst.org/groups/sac/sites/lsst.org.groups.sac/files/LSST2018_datarights_SAC.pdf).

-The plans for reviewing the white papers for LSST cadence optimization. Lynne Jones and Tiago Ribeiro summarized the current status of the Operations Simulator code. (https://project.lsst.org/groups/sac/sites/lsst.org.groups.sac/files/LSST2018_cadence_SAC.pdf).

****Data rights and data access

The draft data rights policy document (http://ls.st/LPM-261), hereafter the DRP, describes what rights to the LSST data mean in terms of how they can be shared and used in publications. We understand that comments on the DRP have been solicited from (and are being received by) a number of constituents, including:

-The international partners, via the LSST Corporation;
-The Science Collaborations;
-LSST Project Personnel.

Here the SAC gives its suggestions and recommendations on this document. First, some general comments:

The DRP attempts to balance the desire to distribute the data as widely as possible with the practical load limitations of the LSST Data Access Centers, and the need to respect the data rights of the US and Chilean communities, as well as those international partners who are contributing to LSST operations costs. The guiding principle of this policy must be to maximize the science opportunities for the LSST data rights community, and thus to remove unnecessary barriers to

getting science done. As we detail below, the SAC felt that there are places where the policy was overly restrictive, in ways that will impede the scientific productivity of the data rights holders. We also felt that the draft rules are overly complex in places, making it challenging to follow them.

In what follows, we develop these concerns and suggestions in detail.

There are many stakeholders in the data rights discussion. As we understand it, this document will ultimately have to be approved by the management of LSST operations (as currently represented by the Interim Management Board), and the National Science Foundation and Department of Energy. This should be made explicit at the beginning of the document.

The document makes a distinction between data rights and data access. As we understand it, those holding data rights are allowed to publish papers with LSST data, while data access specifically allows the individual to access the US and Chilean LSST Data Archive Centers (DACs) via the Science Platform. The document should make the distinction between the two absolutely clear (and if our understanding of the distinction as written here is incomplete or incorrect, this this a reflection of the ambiguity in the current document that will need to be addressed). In an ideal world, we would prefer that this distinction be erased (most other large surveys, including SDSS and Gaia, give full access to all data rights holders), but we understand that limited resources may force the distinction. This distinction leads to many of the complications in the DRP. It will be useful if the document could give examples of individuals with data rights but not data access (i.e., the distinction between “LSST Users” and “LSST Full Users", a confusing terminology in its own right), and how they might interact with the data in practice.

This distinction between data rights and data access is made more complex by the fact that a number of international partners are considering setting up DACs in their home countries. The policies governing these international DACs are being developed by the LSST Corporation; these policies will need to be consistent with the DRP, and the DRP should think through use-cases related to the international DACs. In any case, if enough international DACs are set up, it may be possible to allow data access for all world public data; the international DAC policy should reflect this. Moreover, for the purposes of the DRP document, it should be clarified that "data access" refers specifically to access to the Science Platforms at the US and Chilean DACs.

On a related note, the SAC has long been concerned with the plans to remove data releases more than two years old from the Science Platform. This will make it impossible to reproduce published science results, and will be problematic for people working to complete a science paper based on a specific science release.

The DRP includes rules governing the publication of LSST results, and who may be an author on the resulting science papers. The spirit of these rules are to protect the privileges of the data rights holders, but there are a number of cases in which the rules seem to be overly restrictive, and go against best practices and standards in publication. In particular:

-The definition of "derived data product" (DDP) is unclear, and leads to a lot of confusion. The SAC felt that the current definition (as we understood it) and restrictions on its use and publication would unduly constrain the ability for the LSST data rights community to do science. The DDP concept drives a lot of the implementation complexity in this document. Indeed, other major surveys such as SDSS have not felt the need to define DDP and put restrictions on their publication, and it was not obvious that the DDP is a necessary concept for LSST at all. In any case, if the concept of DDP were kept, it would be useful to give further examples of what would count as DDP and what wouldn't.

-We are quite concerned that the restrictions on publishing LSST data will be in violation of the rules of many scientific journals. One of the examples given in the DRP is the discovery of a dwarf galaxy companion to the Milky Way: while the derived properties of that galaxy could be published, the photometry and positions of the stars that make up that galaxy could not be. It is possible, even likely, that the paper would not be accepted for publication if the authors did not include a table with the LSST data for the stars in the galaxy. And in this case, we saw no harm to the LSST data rights community to having the stellar photometry published. There is a threshold above which it would not be appropriate to publish proprietary data; it would clearly not be appropriate for somebody to publish the detailed photometry of >10^9 LSST galaxies used in a determination of the galaxy luminosity function. But the publication of positions and photometry of 1000 objects may be OK. More thought is needed on where the appropriate threshold should be.

-There is a related loophole that the current rules seem to allow: if a person with LSST data rights were to reprocess the LSST images (using their own software, or say, SExtractor) to measure the properties of the stars in that dwarf galaxy, could they publish that photometry? That would be a somewhat silly way to get around these restrictions, and we were not sure whether

this was consistent with the spirit of the DRP.
-The examples and use-cases given are all for papers in which a single LSST object leads to a discovery, and is the subject of a paper. In practice, only a small subset of papers that come from LSST will fall into this category. For example, what if collaborators without data rights can obtain the spectra of 20 LSST-discovered supernovae? Or have near-infrared photometry for 100 young stellar objects found in LSST photometry? It was unclear whether preventing such collaborations to form and such papers to be written is in the best interest of the LSST data rights community.
-All the use cases describe a situation in which the LSST data are paramount, and the people without data rights add a relatively small component (e.g., follow-up spectroscopy of an exciting LSST-discovered object). But there will be plenty of situations in which the LSST data represent the small increment. Consider, e.g., a particularly interesting transient discovered by the Zwicky Transient Facility, which shows additional activity in the LSST data stream. Would only those ZTF scientists who happen to have LSST data rights be allowed to be co-authors on the discovery paper, if it also includes the LSST light curve? A possibility (which may lead to complications of its own) is to separate the question of data rights from co-authorship, putting in mechanisms to allow non-US/Chilean members of the ZTF collaboration (in this example) to be co-authors on the paper.
-One of the most exciting science opportunities with LSST will be combining the data with other major surveys, such as Euclid, WFIRST, eROSITA, and many others, collaborations which include both LSST data rights holders and those without data rights. The joint analyses will involve not just small samples, but enormous numbers of stars and galaxies. People reading this policy will look for guidance on how such joint analyses might be allowed, and general principles should be articulated here. Given that other large collaborations will have data rights and publication policies of their own, the details will likely need to be worked out in formal Memoranda of Understanding on a case-by-case basis.
-An individual without data rights who is allowed co-authorship on a paper (e.g., to get follow-up spectra of an LSST-discovered object) must be able to see the proprietary LSST data (e.g., a lightcurve) included in that paper. Without this, an author would not be allowed to read the draft of the paper, in violation of basic authorship ethics rules.
-Finally, the DRP states that derived data products are not public until the relevant paper is published in a peer-reviewed journal. This is overly restrictive. For example, this precludes mention of results in Astronomical Telegrams, posting of papers

to ArXiv before they appear in print, or inclusion in any number of unrefereed publications (including SPIE).

The DRP states that it is up to the user to use their own best judgement on matters of data rights. We are concerned that without clearer guidelines, there will be a broad range of interpretations of the rules, causing conflict and resentment in the community, and a general erosion of the rules.

On a related note, the data rights policy must easy to implement and follow, especially if it is applicable to those accessing the data through international DACs. With multiple data releases available at any given time, some public and some not, with people keeping data access for a year after leaving the US, and with no strong enforcement, there will be a lot of inadvertent violations of the rules. LSST needs to put in mechanisms to make it easy for people to follow the rules.

Having a Data Access Policy Committee whose job it is to adjudicate ambiguous cases is a fine idea in principle, but we have several concerns about this.
-Under whose authority would this committee work?
-How would the membership of this committee be chosen?
-Given the significant number of requests this committee is likely to receive, the amount of work this committee would have to do could be quite substantial, equivalent to several FTEs. This will require more than a volunteer effort.

People with LSST data rights and data access include all those who are "US scientists". It was suggested that there would be a master list of US institutions that individuals would have to be affiliated with in order to have data rights; the US DAC is unlikely to have capability for all potentially interested members of the US public. The SAC has several concerns:
-Putting together, updating, and maintaining that list of institutions will be a challenge, to put it mildly. While there would be a mechanism for individuals to suggest additions to that list, we worry that less prominent institutions (for example, community colleges) will be under-represented. It is also unclear how non-educational institutions (including for-profit corporations) at which research is done could or should be included. This gets even less clear for multi-national corporations, even if they are based or centered in the US.
-It is unclear what it means to be affiliated with a US institution. For example, does a US researcher on sabbatical in Paris have data rights and access? A faculty member at NYU Abu Dhabi? A Brazilian

student on a Brazilian fellowship to study in the US? A US citizen working at the University of Tokyo? A Google employee?
-There is also a class of advanced amateurs (in the US and Chile) that will legitimately want access to the Science Platform. We like the idea of having a guide to members of the public, explaining what is available on the EPO site, and explaining the circumstances under which access to the Science Platform is needed. There should be a straightforward mechanism for US residents, unaffiliated with any institution educational or otherwise, to request access to the Science Platform.

The grace period of one year of data rights after leaving the US or Chile to finish a project with LSST data makes sense for senior researchers, but may be overly restrictive for more junior scientists. A few suggestions:
-The policy should be clear that the data rights are for the most recent data release while the individual was still associated with a data rights institution. For example, the individual would not have rights to a new data release that occurs six months after they move outside the US. Moreover, the understanding (perhaps unenforceable) is that these extended data rights would be to finish the project the individual started while at the US, rather than to start new LSST projects at their new institution.
-To encourage the publication of theses based on LSST data, data rights should be extended for US, Chilean, and international affiliate PhD students who leave those countries or institutions, until the end of their first postdoc.

The LSST Project is also developing plans for community event brokers. The issues of data rights and data access are issues for those as well, and we urge that this policy document include use cases both for the 60-second alerts (world-public) and 24-hour releases of nightly data (only available to those with data access).

The SAC did not have time to discuss the thorny issue of who gets access to commissioning data, and how these data would be published. It will be important to work through specific use-cases to come up with reasonable policies. A possible guiding principle is that those who have access to commissioning data should be limited to publishing technical papers until the data become available to the full LSST data rights community.

The SAC looks forward to working with the LSST Project and Operations Team as these draft policies become further refined.

*****Reviewing the Cadence Optimization White Papers

A call has gone to the community to produce white papers for LSST cadence and deep drilling decisions. These are due on November 30, 2018; an informal poll during LSST2018 identified at least 30 white papers in progress (see the growing list at https://community.lsst.org/t/lets-coordinate-observing-cadence-white-papers/3144). In parallel with this, the Operations Simulator team is moving towards the development of a so-called Feature-Based Scheduler, a software architecture that should allow for more flexibility in simulating more sophisticated and complex cadences. This should be available about the time the white papers are due.

With this in mind, the SAC will review the submitted white papers, and suggest a next round of OpSim experiments addressing the various cadence, mini-survey and deep drilling proposals that they suggest. The SAC will produce a first round of feedback to the white paper authors at that time. It will also be the responsibility of the SAC to respect the four science themes (Constraining Dark Energy and Dark Matter; Taking an Inventory of the Solar System; Exploring the Transient Optical Sky; and Mapping the Milky Way) as described in the LSST Science Requirements Document, and to identify any major science opportunities not advocated in the submitted white papers. We expect that few serious suggestions made in the white papers will be rejected at this stage.

A single 10-year run of OpSim takes 50-60 hours of wall clock time to simulate and evaluate, so completing all the OpSim experiments will require of order 8 months. Once these are done, the SAC will review the outputs and the results of the Metric Analysis Framework, and make a final series of recommendations to the Project Office and the Operations Office by the end of 2019.

We plan to incorporate at least a summary of the white papers into the next version of the Community Observing Strategy Evaluation White Paper (COSEP; Marshall et al; https://arxiv.org/abs/1708.04058).