DAP workshop Nov 2019 afternoon notes

https://indico.cern.ch/event/858039/

Resource & service needs

ALICE:

Storage resources for run 2 estimated on 100s of terabytes. Still to discuss within collaboration how many MCs will be included. For Run 3, 2021-2022-2023, more space will be needed (10% of 10 PB AOD per year). AOD format evolution to be determined in the year ahead (May?); applied to new data, not re-filtering of old data.

What is lacking at the moment is manpower. Tested REANA, Open Data Portal successfully. For CAP tests are still in progress, pending configuration issues.

ATLAS:

Current needs are minimal as open data are for education and outreach. For Run 2 the next batch is around 100GB, but unclear on how size will evolve. ATLAS undergoing change to analysis format to a pre-calibrated nano-format. Working assumption is that this will be the next format, 10KB per event (20 billion events). Discussed at the level of the collaboration board: might take up to a year.

(If data is going to increase to multi-petabytes, resource issues must be discussed; 1PB as the threshold for resource discussion).

CMS:
Next CMS release is preliminarily planned for 2020 with slightly more than 1PB (Heavy Ion, 2011 AOD, and first Run2 data)
Publish MC or just code to regenerate? 2011 published all, 2012 categorized and thus selectively published (all available on demand - no demands so far)
To be evaluated whether the cost of all MC is justified. Regenerating is also possible with the available information including the conditions data which are released together with data and MC, but resource intensive for external users.
Open Data is linked intimately to preservation, so if the analysis was done on miniAOD that is what is released
Cost implications of use of nano- / mini- formats; the volume is estimated to be at 1PB/year level even for Run2 as Run2 data format is smaller than Run1 AOD.
CMS does not have the resources to curate smaller datasets for public consumption; preference is to release all verified data after a certain amount of time (following approval of collaboration board)

Data preservation and open data to not be conflated/confused in resource requests

LHCb:

Their data can be quantified in 10s of TBs. For Run 1, the selection output would be a couple of hundred TBs, Run 2 up to a petabyte. Undefined for the coming 3 years.

On the service side, everyone is using open data portal with slightly different uses based on target community. Everyone has been evaluating REANA: ATLAS has clear plan for use of REANA moving forward. For adoption, request base-level resources from IT at the outset as existing REANA installation would need to be expanded.

Question is should resources to support services like REANA come from IT or from the Experiments?

Will this become a mainstream analysis activity or an additional activity? And the resource issues associated with this question

We have to remember that practices that encourage preservation might not be considered a top priority of researchers. But is more useful/efficient for the discipline; so how do we inculcate this change in practice? Practice needs to be incentivized

At the moment, resources are not the primary issue, but adoption/acceptance among the community is likely the most important barrier

Are there some aspects that are experiments responsibility? Content/ experiment specific interfaces. But services could/should be common; customizable to a level that allow their local customization for experiments. It should be scaled to the level that it can be tested for multiple analyses, beyond the existing examples/prototypes. Suggests that CAP might be useful for CERN to place more resources towards it.

All LHC experiments have demonstrated their willingness to integrate their work with CAP. Community should collaborate on evolution of the tool

Funding strategy & risks

Funding agencies will expect data management plans.
This expectation may not be associated with additional resources. Over a certain percentage, a demonstration of value would need to be provided to justify additional investments
Articulate real costs of what we do today, and what implications open data requirements would have on budgets
Research data management plans will require clear articulation of what data will be preserved/opened; funding will require case by case discussions.
RRBs want to know what the experiment policies are on open data and preservation. Common policy may be difficult, but a common statement may be. For example, UK policy that implicates responsibility of hosting lab as opposed to experiment

What we need a common policy on is where the responsibility lies, and what the expectations of the community are. Details of how it is mechanically done can be the purview of the experiments.

Important to understand what is CERNs responsibility, and the experiments' responsibility

Is it the host lab's responsibility to preserve data collected at the experiments and for how long? The life of the experiment? What are the long-term archiving responsibilities / transfer of ownership implications; and associated resource questions for costs/hosting for open data. Who is responsible for the data set once the responsible entity has no resources? Data management plans in other disciplines have defined arrangements for 10 years (sometimes with an external party).

There needs to be a high level policy framework under which the data management plans for the experiments would fit.

This form of policy framework should be ultimately be presented to the RRB for resourcing; following review by the LHCCs

CERN as a host lab should also have a policy around what it hosts, and it should align with the policy framework that encompasses data management practices of the experiments

We should review this as a combined responsibility of the WLCG sites. We can identify the timeframe to be the period of the experiments at the outset, and then modify as needed moving forward

Components that should be host lab / Components that should be WLCG:
CERN could host the services, data could be maintained by the experiments.

Role of hostlab & WLCG sites

- components which have to be host lab? should be?
- data hosting? in the data management plans
- archive can be and is distributed
- CERN can host services, storage can be distributed
- if no expectation of instant access (apart from small samples), perhaps we have the infrastructure we need already
    - overlay it on the infrastructure that we have
    - delivered on request and on some schedule
- CERN resource management will look at how often data is downloaded, not how often it is used
- for open data, other opportunities for storage can arise (e.g. google)
    - some other communities have this model
    - not necessarily free, or something you can rely on in a preservation policy
- different life cycle reqmts: data (open or not) associated with a publication cannot be deleted/obsoleted (at least the ntuple level data - ATLAS policy)
    - reprocessing doesn't invalidate/replace the data associated with the publication
    - there can be convergence between open data and production data requirements, e.g. definitive last reprocessing of a Run-N sample
- Tier expectations
    - interesting to associate the tiers to the open data once it gets big
    - ensure grid resources can be used (via containers). Being pushed by every experiment now. Singularity support everywhere is a request to WLCG.
    -

Common activities & policies

- technical activities:
    - reana/CAP
        - (didn't hear it… someone else please add :-) (missed it too) I think it was something related to acknowledging the contribution from CERN IT/Computing services to support the work done at experiments
- Communication across experiments: what is the best format? Product by product? HSF could be a potential forum for analysis preservation strategies
- WLCG/HSF workshop could include data analysis/preservation (frequency of meetings is every 9 months). Include non data preservation experts, but also distributed computing, etc.
- Notable that everyone is looking at the same set of tools. Important.
- A WG in the HSF as a clearing house for tool discussions, requests visible to all
- policy:
- at a high level clarify a framework agreed by the experiment management

- ideal: a CERN open data policy matching all the experiment policies. Eckhard wants this.
    - Directorate and experiment management will put it on the agenda in one of their monthly meetings
- what to say to the overview board on Thu?
    - Directorate and experiment management will address common policy
    - Commonality on data management plan? Harder. Can't be settled at spokesperson level. Address open data policy first, then data management. They are coupled through their resource requirements
    - all but ATLAS agree on data release in 10 years. (ATLAS doesn't disagree with this, just doesn't have a policy yet)
    - may do reprocessing after 10+ years. So can't necessarily just put the data out there at 10 years.
        - could be a version 2
        - happens with the astro people: data re-issued
    - useful to address 2 phases, they are distinct: live experiment and completed experiment
- would be useful for this group to sketch out a skeleton of the policy. Point out the main parameters
    - same tools
    - everyone at an early stage in trying them out
    - commitment to data preservation is common. Formal agreement expected in the next few years
    - open data has good overlap on reqmts with existing WLCG tier archive infrastructure
    - as much commonality as possible on CERN, experiment data management plans
- RRB visibility of open data: not up to now, but we're going above threshold. Should be mentioned and scrutinized
    - discuss at the overview board, get guidance for 2021