Long-Term Data Preservation: Debriefing Following RDA-4

WLCG GDB, October 2014

Jamie.Shiers@cern.ch

Why?

- Over the past few years, the Research Data Alliance – co-funded by the EU, US and ANDS – has become increasingly important to all things data (and sharing in particular)

- We can use it simply as a “knowledge resource” – **but** also as a way to get funds

- The former is ~~guaranteed~~ has already happened, the latter requires investment (**work**)

➢ **I will explain how…**

DPHEP Background

- The DPHEP Blueprint refers to 4 “levels” of data: http://arxiv.org/pdf/1205.4667.pdf (each with an associated Use Case)
  - Somewhat confusing, conflicts with terminology used by other disciplines and not very accurate
- **Increasingly, we talk (only) in terms of Use Cases, which is:**
  a) More specific;
  b) Matches closely FA requirements (next).
- Open Access (specific samples for outreach etc.)
- Reproducibility of Analyses

➢ **Need for (concrete) Data Management plans**

Data Preservation Levels

Preservation Model	Use Case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

- ***Different preservation models can be organised in levels of increased complexity***
- ***Each level is associated with one or more use cases.***
- ***… it is expected that the cost of various preservation models is primarily driven by person-power requirements rather than the cost of data storage.***

http://science.energy.gov/funding-opportunities/digital-data-management/

- *“The focus of this statement is sharing and preservation of digital research data”*

- **All proposals submitted to the Office of Science (after 1 October 2014) for research funding must include a Data Management Plan (DMP) that addresses the following requirements:**

1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

   If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4).

   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

DPHEP Background

- The DPHEP Blueprint refers to 4 “levels” of data: http://arxiv.org/pdf/1205.4667.pdf (each with an associated Use Case)
  - Somewhat confusing, conflicts with terminology used by other disciplines and not very accurate
- **Increasingly, we talk (only) in terms of Use Cases, which is:**
  a) More specific;
  b) Matches closely FA requirements
- Open Access (specific samples for outreach etc.)
- Reproducibility of Analyses

➢ **Need for (concrete) Data Management plans**

Research Data Alliance: RDA

- Holds 2 plenaries per year, plus short workshops focussing on outputs of Working Groups (WGs)
  - Feb 2014 in Garching; Nov 2014 nr Washington DC; Jun 2015 @ KIT …
- WGs should “complete” in 12 – 18 months – @RDA-4 the first 4 WGs presented their results
  - Next plenaries: March in San Diego, Sep in Japan?
- **On-going debate on value of WGs vs Interest Gs**
  - IGs are longer lived & have less well-defined outputs
- **But, for many, IGs have equal, if not greater, value**
- **E.g. examples of IGs leading to H2020 projects**

RDA – DP Intersection

- "Data Preservation" mentioned in ~every P4 talk
  - "5% cost" discussed repeatedly **("stewardship")**
- Data integrity & preservation were by far the top 2 requirements from sites from survey by "Practical Policy" WG
- Strong interest / support from FAs
- **IGs: preservation, "domain repositories" (merge?)**
- **New IGs: Reproducibility, "Data Fabric", Active Data Management**
- **Certification IG: CTRUST H2020 proposal (4 year)**
  - **Align certification "standards", certify 60+ new sites**

But Also: Co-located Events

- EUDAT, Joint DP workshop, ODIN, DSA, EGI, RECODE, APARSEN, 4C, etc etc
- An excellent opportunity for networking
- Yes, a 5-6 day event is tiring, but less so than 3-4 separate 2 day events with travel

Reproducibility IG

- Can we match the success of the Certification IG and influence future H2020 (and other) calls?
  - https://rd-alliance.org/group/reproducibility-ig.html
- IMHO, if "the RDA" could achieve this, then it would be a highly tangible output and really justify the investment(s)

➢ **We should engage with this group and try to steer it in the right direction**

- **Requires involvement from experiments**
- (Workshop proposed for RDA5)

The Story So Far…

- Together, we have reached the point where a generic, multi-disciplinary, scalable e-i/s for LTDP is achievable – and will hopefully be funded ☺

- Built on standards, certified via agreed procedures, using the “Cream of DP services”

- In parallel, Business Cases and Cost Models are increasingly understood, working closely with Projects, Communities and Funding Agencies

Posit

- Some of us believe that it is possible to analyse the Use Cases of key communities;
- De-compose them into sub-services;
- Provide (at least some of these) via generic tools;
- Whilst at the same time supporting VREs that match the individual / specific requirements of different communities

Why Not at Infrastructure Level?

- Because there really are differences between communities
- Attempting to put “too much” in a “generic infrastructure” has had problems in the past

✓ Equally, we have seen solutions from one community being adopted by others

- A fine balance but let us learn from the past…

VRE Proposal / IG

1. **Prepare a multi-disciplinary proposal to EINFRA-9-2015 attempting to address key Use Cases with a combination of generic services**
   - Matches the call well, which is likely to be heavily over-subscribed
   - EU-JRC interested in this topic
2. **Propose a VRE IG, addressing longer-term issues – targeting a dedicated(?) call in 2-3 years time**
   - This could be more inclusive than the small number of disciplines / Use Cases that could be addressed in EINFRA-9
   - But the number of IGs is mushrooming and the effort to participate is not…
   - **Given that at least some people will not be able to make San Diego, could also submit as a BoF for iDCC 2015 (Feb, London)**

- **How much effort should we invest in "short-term" wrt longer term – more ambitious goals – such as Open Data?**

December pre-GDB

- Given the convergence(?) of at least the LHC experiments on 2 key Use Cases, important to understand what services / support / resources are required / need to be deployed
  - Both from **CERN-IT** and other **WLCG** sites, as well as other projects (**HepData, RIVET, RECAST**, etc.)
  - In particular, what (storage & other) resources are required for Open Access for Outreach?
    - **CDN use case *par excellence?***
    - **An area where also Tier2s could contribute?**
      - Potentially closer to the users?

Summary RDA

- **There are an increasing number of RDA IGs that are of relevance to on-going DP efforts**

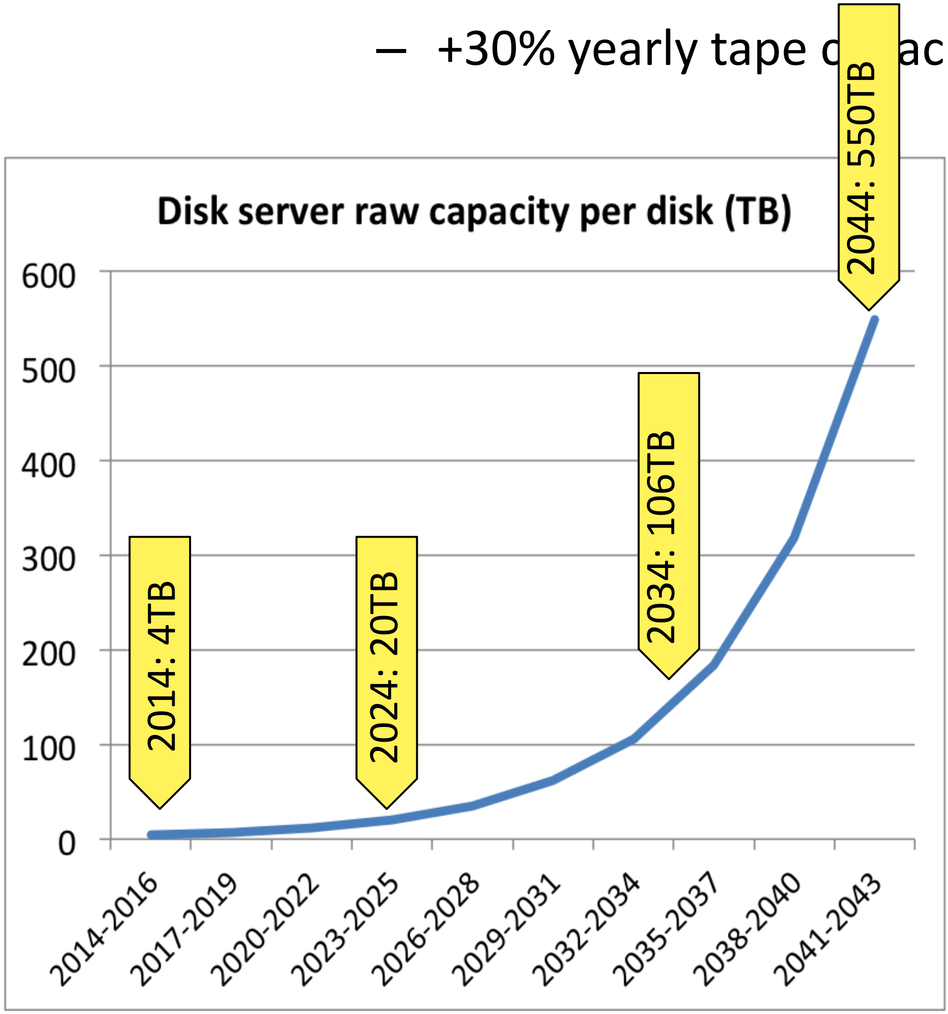
- **At least one has led to an H2020 proposal – opportunity for more**

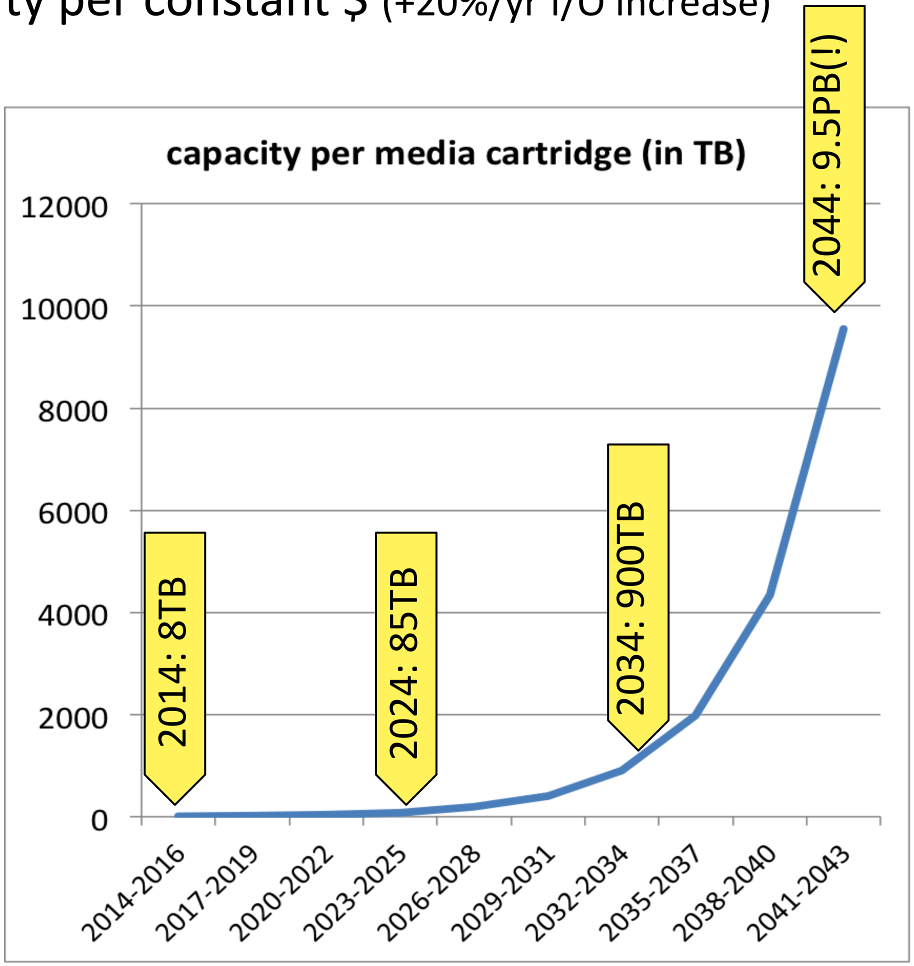
- Together with other projects, a “common vision” not only on the technical aspects, but also on funding (business cases, cost models) & sustainability is being developed

- **The “collective wisdom” that is available at the RDA is impressive – I continue to believe that this is an excellent source of information / knowledge that helps us in a measureable way**

Technology evolution

- Assuming
  - +20% yearly disk capacity per constant $
  - +30% yearly tape capacity per constant $ (+20%/yr I/O increase)





Examples

- LEP: ~100TB = O(10) today’s cartridges

- HERA: ~10PB = O(10) “2030” cartridges

- LHC Run 1: ~100PB =O(10) “2040” cartridges

- ***LHC total: ~10EB = O(??) ????***

Summary DP

- ✓ We are now well known to other data preservation projects & efforts
- ✓ Our (unique?) areas of expertise are respected, as are our cost calculations
- ✓ Convergence on key Use Cases can help to clarify further:
  - – Services, support and resources needed
  - – Opportunities for joint projects / funding
  - + Align / combine efforts with related work (outreach)

2020 Vision for LT DP in HEP

- *Long-term – e.g. FCC timescales*: ***disruptive change***
  - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to annotate further
  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
  - **DPHEP portal**, through which data / tools accessed

➢ **Agree with Funding Agencies clear targets & metrics**

Questions

- **Can we collaborate together on Data Management plans?**

- **Can we work with relevant RDA groups on: Data Sharing / Outreach; Reproducibility; Active Data Management?**

- **Can we prepare for a VRE project whilst (p)reserving enough effort for a “dedicated call”, e.g. on Open Data?**

Conclusions

- Working with / through the RDA and other projects, we are able to establish a "common vision" and inform / influence FAs
- Numerous existing working / interest groups of direct relevance – **more participation would help**
- **Plus also H2020 (and other?) projects**
- **Can take this further:**
  - At the infrastructure level;
  - At the VRE level:
  - Via more ambitious steps, e.g. "Open Data"
  - ➢ **More participation essential**