Executive Summary

- Significant progress has been made in the past years regarding our understanding of, and implementation of services and solutions for, long-term data preservation for future re-use;

- **However, continued investment in data preservation is needed: without this the data will soon become unusable or indeed lost (as history has told us all too many times);**

- **Some of this investment can be done centrally, e.g. by providing bit preservation services for multiple experiments at a given laboratory, whilst important elements need to be addressed on an experiment-by-experiment basis.**

- Funding agencies – and indeed the general public – are now understanding the need for preservation and sharing of "data" (which typically includes significant metadata, software and "knowledge") with requirements on data management plans, preservation of data, reproducibility of results and sharing of data and results becoming increasingly important and in some cases mandatory;

- The "business case" for data preservation in scientific, educational and cultural as well as financial terms is increasingly well understood: funding beyond (or outside) the standard lifetime of projects is required to ensure this preservation;

- A well-established model for data preservation exists – the Open Archival Information System (OAIS). Whilst developed primarily in the Space Data Community, it has since been adopted by all most all disciplines – ranging from Science to Humanities and Digital Cultural Heritage – and provides useful terminology and guidance that has proven applicable also to HEP;

- **The main message – from Past and Present Circular Colliders to Future ones – is that it is never early to consider data preservation: early planning is likely to result in cost savings that may be significant. Furthermore, resources (and budget) beyond the data-taking lifetime of the projects must be foreseen from the beginning.**

Changes With Respect to the Blueprint

With respect to the DPHEP Blueprint, the following observations can be made:

- The pervasive use of INSPIREHEP and other Invenio-based solutions will come as no surprise;
- "Bit Preservation" (and loss) is more clearly defined, with extensive practical experience, albeit different implementations due to site preferences and requirements (hardware choices, funding schemes etc.);

- Virtualisation is more prominent with a better defined timeline (circa ten years);
- The use of CVMFS is a clear success story;
- Cost models and business cases are better understood, with quantitative measures across a variety of experiments;
- "Open Access" policies, embargo periods and the like are new but match well with the *"Zeitgeist"*;
- A variety of "end-of-life" scenarios have been realised: moving from experiment to site support, from host institution to former collaboration members and even porting to new systems and services, such as EUDAT.

These developments, as well as the concrete experience over the past three years, positions the DPHEP Collaboration well to make clear recommendations to future projects and experiments.

Lessons for Future Circular Colliders / Experiments

The main message – from Past and Present Circular Colliders to Future ones – is that it is never early to consider data preservation: early planning is likely to result in cost savings that may be significant. Furthermore, resources (and budget) beyond the data-taking lifetime of the projects must be foreseen from the beginning.

Beyond that, the activities of numerous data preservation activities worldwide can be used as a guide to the type of activities, services and support that is required.

In other words, at least "observer status" from the FCC activities in the DPHEP Collaboration is to be strongly recommended.

For other future and / or current experiments the recommendations are similar:

- Align yourselves with the overall strategy and even implementation of other data preservation activities at your institute / laboratory or globally;
- Adopt mainstream and supported technologies where-ever possible;
- Understand the target communities for your data preservation activities, the Use Cases and the expect benefits and outcomes;
- Try to understand the costs – in particular those that are specific to your collaboration (and not "external" – e.g. host laboratory bit preservation services);
- Data preservation services and support for the LHC experiments can be expected to be provided for several decades: this may be a good place to start.

Future Activities

Over the next period, one can expect progress to be made in the following areas:

- The establishment of a formal policy regarding data preservation for CERN experiments (perhaps linked to the approval process through the Research Board);

- At least a “self-audit” for the CERN Tier0 and WLCG Tier1 sites in the context of the WLCG project;
- Further developments in terms of Analysis Capture and Preservation;
- Further releases of Open Data through the CERN Open Data Portal;
- Harmonization of similar activities across various laboratories and projects;
- Extension of DPHEP’s activities to consider also those of potential FCCs;
- Clarifications regarding funding – of particular importance to past experiments where resources have already become sub-optimal;
- The continuation of regular meetings and workshops, aligning as much as possible with related events (WLCG, CHEP, HEP Software Foundation etc.);
- Further input to the next round of ESPP – building on concrete experience, results and remaining challenges.

The long-term management of the Collaboration also has to be considered – up to 2020 but also beyond.

Outlook and Conclusions

There are clearly many similarities in the approaches being taken, the technologies deployed and the issues encountered. Regular reporting of results (possibly synchronised with major events such as CHEP) should be sufficient to ensure that coordinated approaches remain and that duplication is minimised.

The following quote[1] is traditionally attributed to Leslie Lamport – the initial author of LaTeX and an expert on distributed computing systems.

> A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

This reminds us that data preservation is inherently unstable – with many components and dependencies, constant attention is required to ensure that the entire “system” remains usable. Some changes may be relatively minor, such as a name change in a webserver. Others can be much more disruptive, such as major change in operating system (think VAX/VMS to Unix) or programming language – even a standard-conforming language changes over time, with some constructs being first deprecated, then obsolete and finally unsupported.

Given the cost of today’s storage and the likely evolution, there is no inherent cost why “data” cannot be stored more or less indefinitely. What is harder is to capture the necessary knowledge and validation procedures so that it can be used over long periods of time.

The “natural periodicity” of recent collider generations – some twenty years – is perhaps all one can hope for in terms of affordable data preservation. (Most LEP data – that of ALEPH, DELPHI and OPAL – may be usable somewhat longer, perhaps up to 25 / 30 years). Beyond that, re-use of the data will probably still be possible but

[1] See http://research.microsoft.com/en-us/um/people/lamport/pubs/distributed-system.txt.

may require a larger investment to “resuscitate”, as has been done on rare (one?) occasion(s), notably for the JADE[2] experiment at the PETRA storage ring in DESY.

[2] See https://wwwjade.mpp.mpg.de/ and the DPHEP Blueprint for further information.