2014 – a Watershed[1] Year for DPHEP

Executive Summary

2014 was marked by a number of key achievements in the area of Data Preservation in High Energy Physics. Perhaps the most "famous" of these was the first public release of HEP data, marked by a CERN Press Release at the end of November[2]. However, there were a number of other events that, in time, may be those that have at least as much lasting impact. Firstly, at the beginning of the year, following a topical DPHEP workshop[3] on the "Full Costs of Curation", sufficiently good understanding of the costs were arrived at to make both a medium and long-term plan (more details are given below). This was fed into the relevant planning and budgeting committees and the key resources were agreed. A small fly in the ointment concerns those resources needed for public data but this is a topic that we shall return to later. Throughout the year, many presentations were made to the wider community – via joint workshops with other disciplines, invited talks and panels at policy making / recommending conferences and so forth. This placed Data Preservation in HEP firmly on the map – something that had been previously missing. The year ended with some seven institutes signing the "DPHEP Collaboration Agreement" – key to implementing common strategies and solutions across the world's major laboratories and institutes – as well as agreement between the four major LHC experiments on the Use Cases that drive the need for long-term data preservation. These Use Cases map closely to requirements from major national and international funding agencies, some of which made the provision of a "Data Management Plan" addressing these requirements a pre-requisite for future and repeat funding – again as from 2014.

Introduction to Data Preservation

The theory and practice of data preservation has been driven by disciplines such as Earth Observation and Astronomy that make "observations" – by definition unrepeatable. These disciplines have developed a number of ISO standards that are adopted by almost all communities, worldwide. There is also a significant amount of training material that is freely available based on this set of closely related standards.

Other disciplines – such as public or private archives – have clearly defined policies, often backed by legal requirements. These concern things that might seem mundane, such as tax documents, bank statements, official certificates and other "records". The retention period for such "records" maybe a few decades,

[1] As in "a critical point that marks a division or a change of course; a turning point".

[2] See http://home.web.cern.ch/about/updates/2014/11/cern-makes-public-first-data-lhc-experiments.

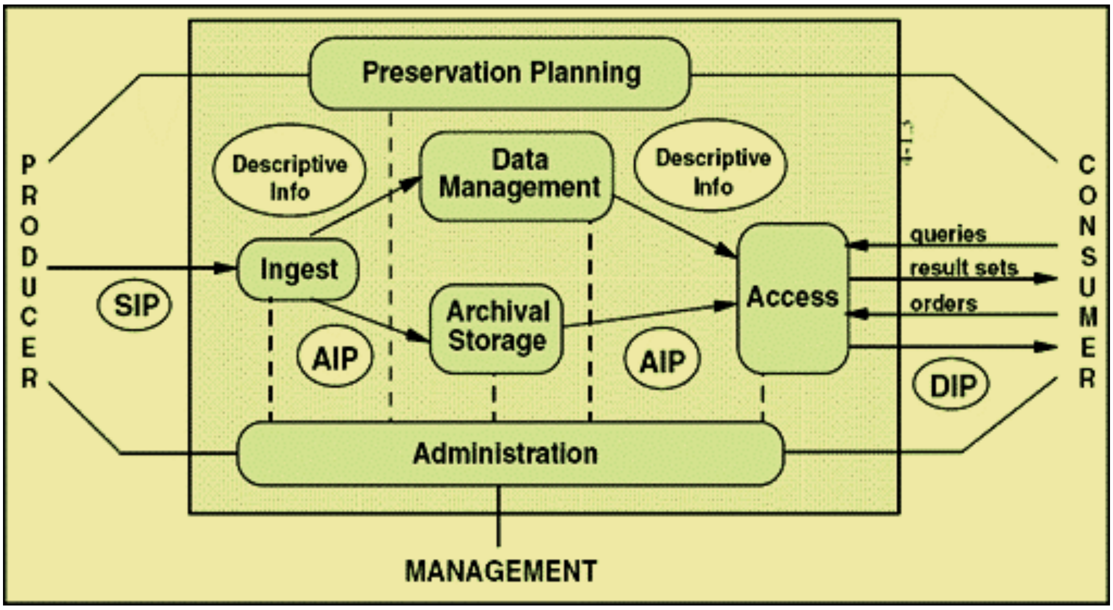
[3] See https://indico.cern.ch/event/276820/ for the full agenda and presentations.

and the formats are typically "simple" (although forward conversion and/or backward compatibility or "emulation" can bring their own problems).

In a domain such as High Energy Physics, the problems are somewhat different. We do not suffer from the "uniqueness" problem of "observations" but nevertheless data from today's and/or recent colliders and accelerators is likely to remain "unique" for a very long period to come. This problem is described in great detail in the DPHEP Blueprint document, available from http://www.dphep.org/. Furthermore, our data are characterized by extremely large data volumes, high data rates and long project lifetimes (although other disciplines are catching up and may well overtake us soon). Whereas others have managed to agree on standard data formats, no such standard exists in HEP. The so-called "standards" that do exist in our community are really closer to I/O libraries, where the I/O and other key information is spread throughout the offline code of the experiment concerned. This is a problem that we will have to face if we wish to address the long-term (re-)usability of data beyond the active lifetime of individuals: the bits may remain but the ability to interpret them may well be lost.

Overview of the OAIS Reference Model and Related Standards

The Open Archive Information System (OAIS) is a reference model that originated in the Space community. It was originally defined by the Consultative Committee for Space Data Systems (CCDS) in 2002, revised in 2012, and is available for download from their website[4] and/or for purchase from the International Standards Organisation ISO), where it is known as ISO 14721:2003.



An important feature of this reference model, shown above, is that it differentiates between the *producers* and *consumers* of the information. Depending on the knowledge base of the latter, varying amounts of information and/or metadata need to be archived with the data itself, requiring one to think carefully ahead.

[4] See public.ccsds.org/publications/archive/650x0m2.pdf.

The reference model is now ubiquitous and a wealth of training material can be found on the Web[5] and/or in manuals and even books. OAIS is not without its critics but the fact that it is so widely accepted makes it a valuable tool to facilitate discussions between disparate communities.
Building on top of OAIS are a number of certification procedures, including ISO 16363[6] and ISO 16919, which specifies the competencies of the audit team. According to the Primary Trustworthy Digital Repository Authorisation Body (ISO-PTAB) *"ISO 16919 is vital to enable an audit process which is performed with impartiality, competence, responsibility, openness, confidentiality and responsiveness to complaints, supported by the international ISO process."* PTAB runs a series of training courses on all of these standards and one such course is foreseen to be held at CERN, primarily for the WLCG Tier0 and Tier1 sites, during 2015.
Full details of these standards and the associated training materials is outside the scope of this paper. However, a simpler certification procedure exists, known as the Data Seal of Approval[7] (DSA), developed by the Data Archiving and Networked Services (DANS) institute in the Netherlands.
DSA defines a total of 16 guidelines and offers a "self-assessment", as well as more formal auditing options. Of these guidelines, 3 concern the producer, 3 the consumer and the remaining 10 the repository itself. These are reproduced below.

Guidelines Relating to Data Producers:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.

2. The data producer provides the data in formats recommended by the data repository.

3. The data producer provides the data together with the metadata requested by the data repository.

Guidelines Related to Repositories:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

[5] See, for example, http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/using-oais-reference-model-curation, from the Digital Curation Centre in the UK.
[6] Also available for free via public.**ccsds**.org/publications/archive/652x0m1.pdf.
[7] Within the context of the Research Data Alliance (RDA), there is ongoing work to ensure the harmonization / compatibility of the various certification procedures. There is also agreement amongst the "owners" of the procedures that they can be viewed as a hierarchy, with DSA as the "entry level".

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

6. The data repository applies documented processes and procedures for managing data storage.

7. The data repository has a plan for long-term preservation of its digital assets

8. Archiving takes place according to explicit work flows across the data life cycle.

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

10. The data repository enables the users to discover and use the data and refer to them in a persistent way.

11. The data repository ensures the integrity of the digital objects and the metadata.

12. The data repository ensures the authenticity of the digital objects and the metadata.

13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Guidelines Related to Data Consumers:

14. The data consumer complies with access regulations set by the data repository.

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

16. The data consumer respects the applicable licenses of the data repository regarding the use of the data.

Whilst a number of these guidelines might seem simply "common sense", having explicit and auditable practices greatly increases the chances that the data will indeed be preserved – and that it will be usable a number of decades hence. Take, for example, data from previous HEP experiments, such as those at LEP. Whereas the ALEPH experiment used the DESY Bos package to read and write their data, the others (DELPHI, L3 and OPAL) used Zebra from the CERN Programme Library. Not only is this insufficient information to be able to interpret the data in the fully general case (some Zebra "exchange" formats can be understood by first reading the first few bytes of each file) but the original authors are no longer available and – until recently – the documentation had been "lost". (The webserver hosting the CERNLIB "long write-ups" no longer exists – the PS and HTML files were found on a Web archive machine and the sources have since been recovered. There is a project in preparation to archive these write-ups, as well as those from the experiments, in a digital library with extensive metadata in an *attempt* to ensure their persistence).

Equally, guidelines 4, 6, 7 and on are close to mandatory if one is to have *any* guarantee that the data will be available in the long-term.

Furthermore, it is widely understood that having a single copy of the data is insufficient to protect against common causes of data loss. Whereas the LHC experiments maintain a 2nd copy of (most of) their data across the WLCG Tier1 sites, this is not the norm for many other experiments.

This may become a policy for at least CERN experiments in the coming years but it needs to be matched by concrete implementation!

Towards a Data Preservation Strategy for CERN Experiments

The updated Strategy for European Particle Physics[8], approved by Council in May 2014, states that *"infrastructures for … data preservation … should be maintained and further developed."*

In order to implement this strategy, the following proposals are currently under discussion. (The numbering reflects the draft proposal, where the paragraph above is point 1.):

2. Such infrastructures include *digital repositories*, where *copies* or *replicas* of the data are kept.
3. As host laboratory, it is expected that (from now on?) a copy of all data acquired by CERN experiments *and* targeted for long-term preservation be stored in the CERN digital repository. This will typically include all raw data and the final reprocessing pass and associated Monte Carlo datasets.
4. It is strongly recommended that one or more copies of the above data are maintained outside, at or spread over institutes that form part of the collaboration.
5. In order to ensure sufficient reliability and adherence to "best practices", it is recommended that such repositories follow agreed guidelines / standards – this is currently being discussed in the context of WLCG for LHC data.
6. These guidelines not only include policies for the management of the repository itself, but also on access to data in the repository (adherence to agreed access policies and terms of use), as well as the *ingest* process, when

[8] See http://council.web.cern.ch/council/en/EuropeanStrategy/ESParticlePhysics.html.

data is “entered” into the repository. The latter is to ensure that appropriate and supported data formats are used, there is sufficient documentation, meta-data and other materials to permit use by the designated communities, and so forth.

7. The above recommendations could become part of a default strategy for CERN experiments, with implementation details – including variances on the above – provided in the Data Management Plan (DMP) for that experiment. DMPs are increasingly required by funding agencies for new / repeat funding and can be expected to be quasi-mandatory in the future.
8. As a minimum, the DMP of an experiment should detail the policy for storing replicas of data and the recovery mechanisms, both during and after the active lifetime of the associated collaboration.
9. These basic recommendations are expected to be supplemented by others – e.g. on “knowledge capture and preservation” – as we gain experience with preserved and open access data.

It is foreseen that this proposal will be discussed at CERN’s scientific committees – most likely starting with the LHCC, as an implementation based on the WLCG Tier0 and Tier1 sites could be a reality in the short to medium term.

The DPHEP Study Group

The DPHEP study group was initiated in early 2009 and became a sub-group of the International Committee for Future Accelerators (ICFA) – emphasizing its global nature – later that year. Its goal was:

High Energy Physics experiments initiate with this **Study Group**[9] a common reflection on **data persistency and long-term analysis** in order to get a common vision on these issues and create a multi-experiment dynamics for further reference.

The objectives of the Study Group are:

- Review and document the physics objectives of the data persistency in HEP.
- Exchange information concerning the analysis model: abstraction, software, documentation etc. and identify coherence points.
- Address the hardware and software persistency status.
- Review possible funding programs and other related international initiatives.
- Converge to a common set of specifications in a document that will constitute the basis for future collaborations.

As well as running a series of workshops that rotated around all of the main HEP laboratories, it generated a Blueprint document that was well received by ICFA and was fed into the process for updating the European Strategy for Particle Physics.

The full Blueprint – which runs close to 100 pages – should be referred to for details regarding the motivation for and status of data preservation activities across all key laboratories (status in 2012).

[9] See http://dphep.org for further details.

It states:

"Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP. This paper includes and extends the intermediate report. It provides an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organisation in charge of the data management and policies in high-energy physics."

The DPHEP study group identified the following priorities, in order of urgency:

- ***Priority 1: Experiment Level Projects in Data Preservation.*** *Large laboratories should define and establish data preservation projects in order to avoid catastrophic loss of data once major collaborations come to an end. The recent expertise gained during the last three years indicate that an extension of the computing effort within experiments with a person-power of the order of 2-3 FTEs leads to a significant improvement in the ability to move to a long-term data preservation phase. Such initiatives exist already or are being defined in the participating laboratories and are followed attentively by the study group.*

- ***Priority 2: International Organisation DPHEP.*** *The efforts are best exploited by a common organisation at the international level. The installation of this body, to be based on the existing ICFA study group, requires a Project Manager (1 FTE) to be employed as soon as possible. The effort is a joint request of the study group and could be assumed by rotation among the participating laboratories.*

- ***Priority 3: Common R&D projects.*** *Common requirements on data preservation are likely to evolve into inter-experimental R&D projects (three concrete examples are given above, each involving 1-2 dedicated FTE, across several laboratories). The projects will optimise the development effort and have the potential to improve the degree of standardisation in HEP computing in the longer term. Concrete requests will be formulated in common by the experiments to the funding agencies and the activity of these projects will be steered by the DPHEP organisation.*

These priorities could be enacted with a funding model implying synergies from the three regions (Europe, America, Asia) and strong connections with laboratories hosting the data samples.

The DPHEP Collaboration Agreement

In order to implement priority 2 above (experiment-level data preservation is already under way in most cases and common "R&D" projects are already leading to services with a view to long-term support and sustainability), CERN has appointed a Project Manager (October 2012) and a Collaboration Agreement has been prepared. 7 institutes have now signed this agreement (CERN, DESY, HIP Finland, IHEP, IN2P3, KEK and MPP) with several more in the pipeline.

The agreement, which largely reflects the recommendations of the Blueprint, includes the following goals:

The Project, in coordination with the International Committee for Future Accelerators (ICFA), aims at:

1. *Positioning itself as the natural forum for the entire discipline in order to foster discussion, achieve consensus and transfer knowledge in two main areas:*

    a. *Technological challenges in data preservation in HEP,*
    b. *Diverse governance at the collaboration and community level for preserved data,*

2. *Co-ordinate common R&D projects aiming to establish common, discipline-wide preservation tools,*
3. *Harmonize preservation projects across the Partners and liaise with relevant initiatives from other fields,*
4. *Design the long-term organization of sustainable and economic preservation in HEP,*
5. *Outreach within the community and advocacy towards the main stakeholders for the case of preservation in HEP.*

All of these areas are currently being pursued actively and can be viewed in terms of a (slowly evolving) "2020 vision".

The DPHEP 2020 Vision

The "vision" for DPHEP consists of the following key points:

- 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

- There should be a **DPHEP portal**, through which data / tools accessed

- **Clear targets & metrics** to measure the above should be agreed between **Funding Agencies, Service Providers** and the **Experiments (Collaborations).**

Although there is clearly much work still to be done, this vision looks both achievable and the timescale for realizing it has been significantly reduced through interactions with other (non-HEP) projects and communities.

Key LHC Use Cases

The Use Cases that drive data preservation that have been agreed by the four main LHC experiments are as follows:

1. Preserve data, **software, and know-how** in the collaborations (this includes **bit preservation** of the raw data as well as (versions of) the derived products);
2. Share data and associated **software** with larger scientific community;
3. Open access to reduced data sets to general public.

These match closely, but are more explicit than, the requirements coming from funding agencies. It is likely that these will become more detailed with time, particularly in the area of analysis reproducibility (which some differentiate from *replicability*).

Requirements from Funding Agencies

There have been numerous policy discussions and recommendations in recent years, some of which are reflected in the outputs of the (EU FP7) projects discussed below. A particularly clear statement can be found from the US Office of Science[10] that includes the following:

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:

- ***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

  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

[10] See http://science.energy.gov/funding-opportunities/digital-data-management/.

Similar requirements are coming (or have come) from other Funding Agencies and for International projects in particular it will be important to understand how to respond to these in a consistent manner. This is part of the debate that will continue, e.g. following the RECODE project recommendations covered below.

Open Access Policies

The four main LHC experiments have approved Open Access policies[11] that, whilst they differ in detail, are broadly similar (and are being adopted by other experiments):

1. (Moving towards) Gold Open Access for Publications (DPHEP "level 1");
2. Open Access to Specific Data Samples for Outreach (DPHEP "level 2");
3. Open Access to (a fraction of the) Reconstructed data (after an embargo period) (DPHEP "level 3");
4. Raw data[12] closed even to collaboration (today) (DPHEP "level 4").

The "fractions" involved vary from 30-50% after a few years to (in some cases) 100% after ~10 years. (Just under 40TB of CMS data from 2010 have been released and 10TB of ALICE pp and PbPb data are expected to be released shortly. LHCb will release their first data only in 2018. For ATLAS, the plans are still unclear, but a volume similar to CMS or ALICE can be expected).

Even though this applies to the reconstructed data, the volumes involved could end up being very significant and the technical and financial issues, particularly in the medium to long term (2020+) are not yet understood!

4C and RECODE Policy Recommendations

4C was an FP7 project that terminated in January 2015 to help clarify the costs involved in data curation. Its goals were:

"4C will help organisations across Europe to invest more effectively in digital curation and preservation. Research in digital preservation and curation has tended to emphasize the cost and complexity of the task in hand. 4C reminds us that the point of this investment is to realise a benefit, so our research must encompass related concepts such as 'risk', 'value', 'quality' and 'sustainability'."

Its roadmap document[13] contains the following recommendations:

[11] See http://opendata.cern.ch/collection/data-policies.

[12] Most disciplines use a different notation, with "L0" corresponding to the raw data and L1/L2/L3 corresponding to calibrated and/or processed and/or derived data.

[13] See http://4cproject.eu/.

1. *Identify the value of digital assets and make choices;*
2. *Demand and choose more efficient systems;*
3. *Develop scalable services and infrastructure;*
4. *Design digital curation as a sustainable service;*
5. *Make funding dependent on costing digital assets across the whole lifecycle;*
6. *Be collaborative and transparent to drive down costs.*

With its leadership in providing scalable, sustainable services, HEP is well positioned to make key contributions in many of these areas. However, we must be aware of and plan for recommendation 5, which could have significant funding implications!

The Policy RECommendations for Open Access to Research Data in Europe (**RECODE**) project:

"will leverage existing networks, communities and projects to address challenges within the open access and data dissemination and preservation sector and produce policy recommendations for open access to research data based on existing good practice."

As for 4C, this was also an FP7-funded project that recently terminated, again with a final set of policy recommendations.

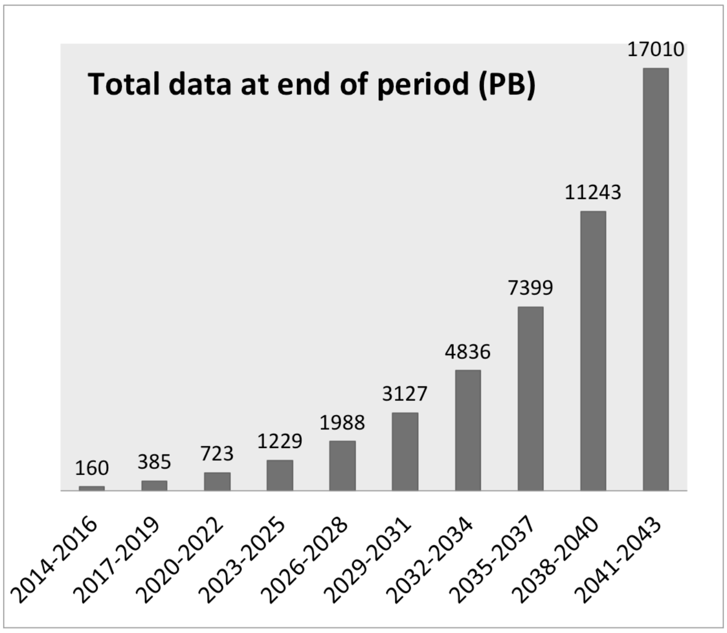
As has happened with publications, the most likely course of events is that the Open Access to data movement will gain momentum. However, given the above-mentioned LHC policies and the volumes of data involved, we need to be prepared to answer the following questions:

1. Is it financially affordable?
2. Is it technically implementable?
3. Is it scientifically (or educationally, or culturally) meaningful?

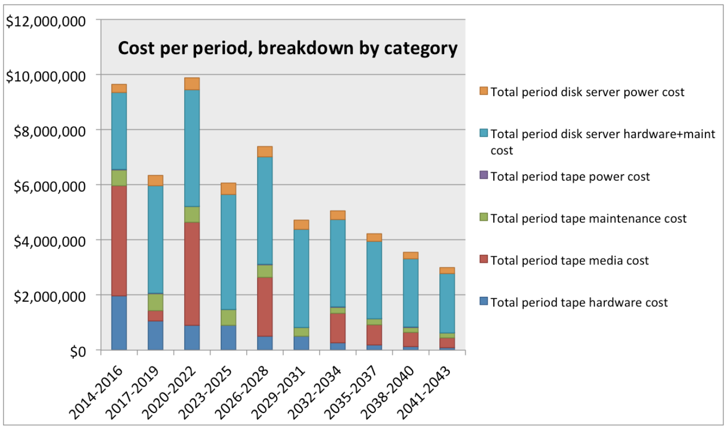
The answers to these questions may well vary with time – as we will see shortly when we consider data from the LEP experiments – and also depend on the implementation(s) that we choose: Open Access is just one step in the progression towards Open Data and finally "Open Knowledge".

Bit Preservation

A very simple model that loosely matches the expected evolution in acquired LHC data volumes is shown below.



Based on publically available technology predictions and pricing information, we are able to calculate how much it would cost to store a single copy this information in a set of tape libraries (a 10% disk cache is included, as is a 3-year cycle for the media, after which all data are migrated forward to the next generation).
Not only does this – together with on-going data scrubbing – implement some of the key practice in OAIS and the associated certification procedures and hence allow us to offer "state-of-the-art" bit preservation, but we can also calculate the costs of a data store rising from several tens of PB initially to a few EB in the 2030s. Whilst much more detailed calculations are used in the LHC (WLCG) budget review and request process, this gives us at least a ballpark estimate for the costs involved and we can see that the cost over time averages to "just" $2M / year (for such a vast and growing data store). Comparatively, e.g. versus the cost of LHC computing, the cost of building and running the machine and its detectors, this is a "small number" – certainly much less than the cost of building a new machine in the future (at least with today's technology)!



Bit preservation is an art in itself and – following the 4C project recommendations – is best performed at a limited number of "expert" sites, rather than across a multitude of smaller ones. This becomes even more important as densities increase – whereas user manipulation of individual tape volumes was common place in the LEP era, the latest generations of media requiring extreme clean-room conditions and prefer robots over humans!

LEP Data

Prior to the LHC – a "discovery machine" – another collider was housed in the same tunnel. This collided electrons and positrons and was a "precision machine" to study the Z and W particles that had been observed in a previous "discovery machine" – the CERN Super Proton Synchrotron operating as a proton anti-proton collider. During the running period of LEP (1989 – 2000), the computational and storage requirements were enormous for their day. However, technology evolution has since made these requirements almost trivial[14] (using today's technology). Indeed, the full LEP dataset of some 100TB could be stored on O(10) of today's latest generation tape cartridges! Using the technology predictions that are behind the storage cost estimates shown above, one can guestimate when the full Tevatron or Hera data will be storable on O(10) "cartridges" (perhaps in the 2030s) or the full LHC Run1 data set (in the 2040s?). Without being too specific, it is probably safe to say that today's daunting challenges become tomorrow's routine: simply "adjusting" the duration of the embargo period and/or adopting a more drawn out release schedule may be enough to resolve the financial and technical challenges listed above.

The LEP data is still valuable and publications based on it are still appearing[15]. Copies of the ALEPH data exist in Italy (INFN) and of the OPAL data in Germany (MPP). Attempts to preserve not only the data but also "the knowledge" are still on-going, benefiting from the latest ideas, technologies and services. Thus, one can expect that the LEP data will still be usable – at least by experts – two decades after the end of data taking, if not three decades or possibly more.

Whereas much of this work was started at the end of LEP data taking or even after, this experience will be invaluable in preparing the long-term preservation of data from the LHC, which still has a very long active life ahead of it.

Data Portals – HEP, CERN and elsewhere

Data portals have existed for a long time but the concept of a Data FAIRport[16] comes from 2014:

"Data FAIRport is an open initiative of representatives from the worlds of research infrastructure and policy, publishing, the semantic web and life sciences research. Data FAIRport was founded in January of 2014. It's vision is to realise and enable a

[14] See the LEP presentations at https://indico.cern.ch/event/276820/other-view?view=standard.

[15] See, for example, this WIRED article about a Higgs analysis based on ALEPH data: http://www.wired.com/2015/01/higgs-discovery-hijack-attempt/.

[16] See http://datafairport.org/.

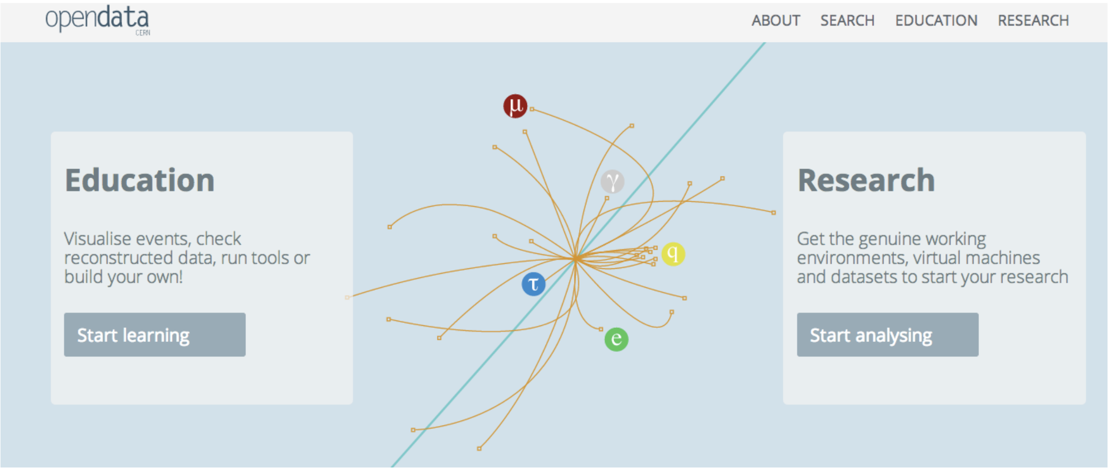
situation where valuable scientific data is ‘FAIR’ in the sense of being Findable, Accessible, Interoperable and Re-usable.”

This is certainly compatible with the DPHEP 2020 vision – with the added emphasis on “interoperable”. For HEP, a minimum first step would be to make the data from two experiments “interoperable” – heading in the direction of Open Data. (What interoperability means between data from HEP and another discipline is less obvious, unless a closely related one such as astro-particle physics. It is more meaningful in disciplines such as Earth Science or Astronomy, where the combination of data from (completely) different “sensors” can have rather immediate value. A striking example is that of the GOCE satellite which

“became the first seismometer in orbit when it detected sound waves from the massive earthquake that hit Japan on 11 March 2011[17]*”*

Numerous other examples exist of data being re-used for purposes sometimes wildly different to those originally foreseen. This indeed is one of the principles of the Research Data Alliance[18] (RDA) that encourages “Research data sharing without barriers”.

The current CERN Open Data portal (see below) is relatively modest in scope and clearly needs to be extended to other experiments[19] at CERN – past, present and future.



Education

Visualise events, check reconstructed data, run tools or build your own!

Research

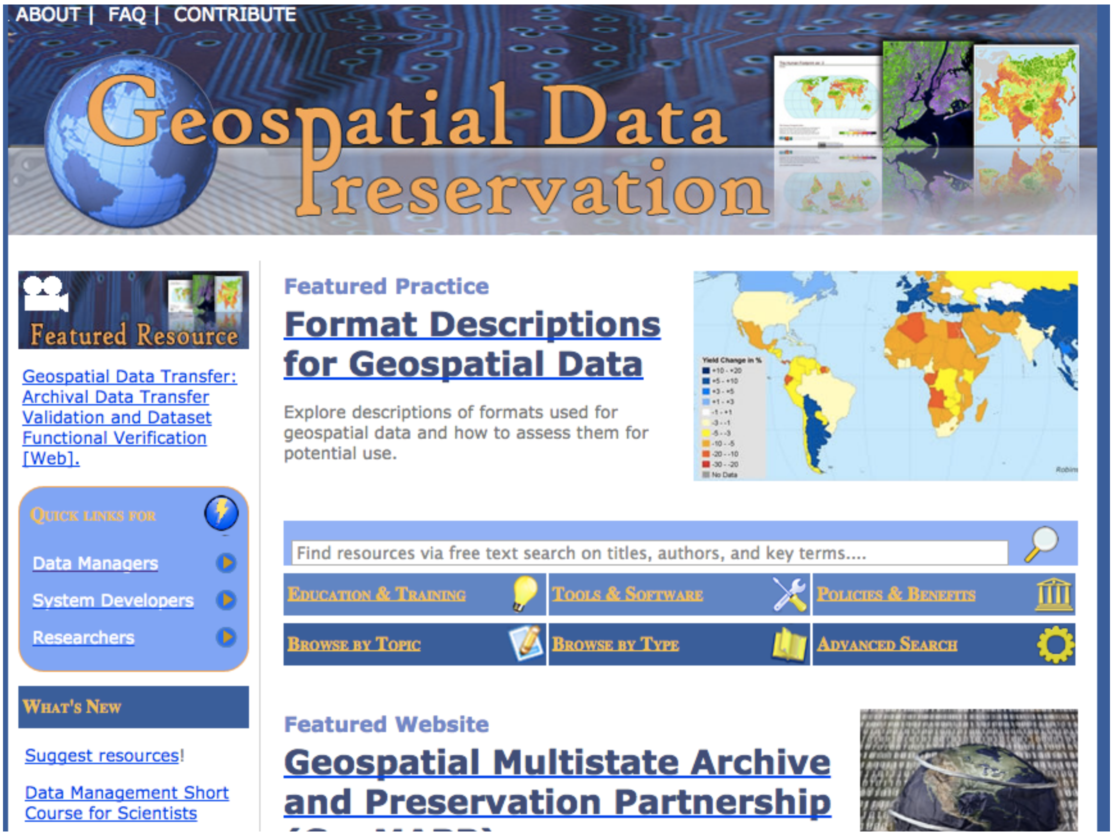
Get the genuine working environments, virtual machines and datasets to start your research

Further extensions include a higher-level portal that allows also access to data from experiments at other HEP laboratories. Beyond that, one could extend further into something like a “knowledge base” that again exists for numerous other disciplines.

[17] See http://www.esa.int/Our_Activities/Observing_the_Earth/GOCE/Surpassing_expectations for further details.

[18] See https://rd-alliance.org/node.

[19] See http://greybook.cern.ch/ for a complete list of such experiments (albeit in a somewhat “terse” format.



ABOUT | FAQ | CONTRIBUTE

Geospatial Data Preservation

Featured Resource

Geospatial Data Transfer: Archival Data Transfer Validation and Dataset Functional Verification [Web].

Quick links for

Data Managers

System Developers

Researchers

What's New

Suggest resources!

Data Management Short Course for Scientists

Featured Practice

Format Descriptions for Geospatial Data

Explore descriptions of formats used for geospatial data and how to assess them for potential use.

Find resources via free text search on titles, authors, and key terms....

Education & Training | Tools & Software | Policies & Benefits

Browse by Topic | Browse by Type | Advanced Search

Featured Website

Geospatial Multistate Archive and Preservation Partnership

Open Data and Beyond – Open Knowledge

Data preservation is in a number of senses an unfortunate term. Firstly, it hides the real motivation, such as the Use Cases we have highlighted above. Secondly, it gives the impression that it is "just" about the data, whereas in disciplines such as HEP the data is only part of the story. To be effectively used, HEP data requires complex algorithms, a great deal of meta-data (such as information about the detector(s) for each event in term), Monte Carlo simulations and so forth. To attempt to go beyond Open Access, beyond Open Data and to include this "knowledge capture" in a meaningful way, a consortium that includes a HEP institutes (CERN and INFN), as well as those from quite different disciplines have put together a project proposal call Open Knowledge Science. This proposal was only submitted in the early days of 2015, having been developed through much of 2014.

The vision is grand: To create, across the globe, the means for researchers to work together with ever-growing data flows – instantly, efficiently and creatively, crossing all boundaries of discipline, geography, institution or policy. In short: To make the whole world a single, living lab. This laboratory of Open Science will have all the functionality required to allow data to be analysed, validated, integrated and re-used. For this to happen, new tools and services must be layered on top of the physical computer networks, servers and storage systems (RDA Europe, 2014, pg. 13).

The vision put forward by the Research Data Alliance Europe in its "Data Harvest" report captures well the challenges ahead of the international scientific community; its call for a (virtual) laboratory of Open Science echoes the characteristics of the Virtual Research Environments (VREs) identified in the call EINFRA-9-2015.

Open Science builds on knowledge sharing for data (re-)usability and the reproducibility of scientific results, a key concern of today's science1**. This goes beyond the current movement for open access to data, it requires open access to knowledge.**

We propose an Open Knowledge VRE for Science focusing on the production and sharing of Knowledge across scientific domains within the Open Science paradigm, pushing the boundaries of today's state of the art.

Conclusions and Outlook

Not all of the tools or services required or even in use to tackle data preservation in HEP have been covered in the above. The glaring omissions are justified by what could reasonably have been covered in a series of 3 lectures as well as those topics the author felt sufficiently competent to address. In particular, the vast amount of work that needs to go on within the Collaborations is clearly missing, as is information on digital library tools, virtualization technologies and so forth. A future, multi-author, version of this paper may address them.

The outlook, however, is encouraging. Not only do we have a reasonable to good understanding of the issues involved, but also of potential – and even practical – solutions, as well as their costs and plans for long-term sustainability.

How well we succeed in tackling the really difficult problems, such as Open Data and Open Knowledge, remain as "Open Questions" for the future.