ILDAP

International Long-term Data and Analysis Preservation

V1.11
18 November 2011

Part B

Type of funding scheme:

Coordination and Support Actions – coordinating actions (CSA-CA)

Work programme topic addressed:

INFRA-2012-3.2 International cooperation with the USA on common e-infrastructure for scientific data.

Name of the coordinating person:

Jamie SHIERS

List of Participants:

Participant number	**Participant organisation name**	**Participant short name**	**Country**
1 (Coordinator)	EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH	CERN	CH
2	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)	CNRS	FR
3	STIFTUNG DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY	DESY	DE

Table of Contents

1. Scientific and technical quality

1.1. ILDAP Concepts and Objectives

Objectives

1. Through collaboration with related projects in the US, take the recommendations of the Data Preservation in HEP for Long-Term Analysis (DPHEP) study group and carry them to a wider scientific scale in terms of scope and the LHC scale in terms of volume and duration;
2. Establish the requirements and standards to be used throughout the project and develop first a demonstrator and later a prototype based on these (see associated milestones and deliverables);
3. Ensure the full engagement of the LHC experiments, themselves international collaborations with significant components in both Europe and the US, in data preservation activities for the data that is currently being acquired.

Motivation for International Cooperation

The scientific and technical reasons for performing data preservation are given below. The motivation for performing this work through international cooperation – and in particular with the US – can be summarised as follows:

- High Energy Physics (HEP) is global by construction: today's facilities are far larger and more expensive than can be afforded by any one country and are now operated as global facilities with participation from institutes in some tens of nations;
- Preservation is a global problem and "owners" of the data set – that is those who paid for and those who acquired them – include a variety of funding agencies: those in Europe, those in the US (including both the Department of Energy (DoE) and the National Science Foundation (NSF)) as well as elsewhere in the world;
- There is active support for the activities described in this proposal both within Europe as well as in the US; complementary projects are being established in the US to work together with ILDAP on this global challenge (see letters of support in Annex 8);
- Hands-on experience with a number of issues in the realm of data preservation exists in projects under the DPHEP umbrella. This includes DESY – a partner in the ILDAP proposal – and also SLAC in the US;
- Platforms, such as INSPIRE, are used for preserving documentation and metadata. INSPIRE is jointly operated by CERN and DESY as well as DoE-funded laboratories, such as SLAC and FNAL.

In summary, only a global response can address what is clearly a global challenge.

Introduction

The preservation of scientific data for long-term use and re-analysis has been identified as a key requirement in the field of High Energy Physics (HEP) and other disciplines such as Astronomy and Astrophysics, as well as Life and Earth Sciences. In collaboration with related projects in the US (in particular in close collaboration with the National Science Foundation and Department of Energy) the proposed project would take the work of the Data Preservation in HEP Study group [DPHEP] that defines the physics motivation for long-term data preservation and many of the associated issues, and extend this to cover not only the existing use cases from facilities which have recently stopped operation but also consider the emerging needs of the LHC experiments at CERN. This work would ensure the persistent availability of existing data and enable it to be shared between organisations and across national boundaries.

Now is the time to define standards for data and meta-data formats and address socially pressing issues such as possible Open Data access in the context of on-going experiments (e.g. those at the LHC) – issues that have historically been addressed only in the final years of a scientific collaboration if at all. A call for action was captured[1] in a large-scale survey of the HEP community performed as part of the PARSE.Insight Support Action, also funded under FP7.

In order to perform this work a coordination body would be established that will organise workshops devoted to this topic and also address key issues related to long-term data archives, such as infrastructure approaches, the financing models for maintaining these archives, and issues of authorization and trust for access to the data both during and after the lifetime of the corresponding scientific collaboration, considering opportunities offered by Open Data concepts weighed against the community practices and self-organising collaborations. Moreover, networking of experts, both within the HEP domain and with other disciplines and projects, will be established.

The results of this work would both be made immediately available via Open Access mechanisms and actively disseminated at relevant technology-oriented events, such as the IEEE Massive Storage and Technology conference, as well as discipline-focussed meetings, such as the IEEE Nuclear Science Symposium and Medical Imaging Conference and other similar events. As HEP is currently a large user of grid computing, contacts in this domain would also be used, such as the EGI Technical and Community Fora in Europe and Open Science Grid meetings in the US.

Long term preservation of HEP data is crucial to preserve the ability of addressing a wide range of scientific challenges and questions at times long after the completion of experiments that collected the data. In many cases, these data are and will continue to be unique in their energy range, process dynamics and experimental techniques. New, improved and refined scientific questions may require a re-analysis of such data sets. Some scientific opportunities for data preservation are summarised below.

[1] See http://arxiv.org/abs/0906.0485 and http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf.

Long-term completion and extension of scientific programmes
This entails the natural continuation of the physics programme of the individual experiments, although at a slower pace, to ensure a full exploitation of the physics potential of the data, at a time when the strength of the collaboration (analyst person-power as well as internal organisation) has diminished. More important than the sheer number of publications is the nature of these additional analyses. Typically, these analyses are the most sophisticated and benefit from the entire statistical power of the data as well as the most precise data reprocessing and control of systematic effects.

Specific issues that data preservation addresses include:

1. Cross-collaboration analyses – the combination of data from multiple experiments to either reduce statistical and/or systematic uncertainties of single experiments, or to permit entirely new analyses that would be otherwise impossible;
2. Data re-use – for example, when called for by new theoretical insights, new or improved analysis and/or simulation techniques, or to search for previously unseen signals observed at a newer facility.

Further details on these, together with concrete examples of the re-analysis of data, are given in annex 5.

In summary, the persistence of data analysis in HEP can and has led to new scientific opportunities. The examples given may well be only the tip of the iceberg. The paradigm itself and the issues addressed (standards, longevity, robustness, cross-collaborations) may have a structural impact in HEP data analysis and may shape the future of the computing models (as it did for astrophysics). At the time when data is increasing rapidly (see astrophysics or life sciences), HEP will need a data analysis approach that is closer to other sciences and may play a leading role in building synergies for future Experimental Data Observatories.

However, whilst HEP has been at the forefront of devising ways to “ride the wave” of the rising tide of scientific data (as exemplified by its use and evangelism of grid computing), it is lagging behind other fields in terms of data preservation. By bringing HEP to the same level – that is as a pioneer and driving force – we would thereby create unique conditions for virtuous data observatories.

Data Preservation and the LHC
The experiments at the LHC are foreseen to continue for at least 20 years based on the present schedule of the LHC project. There is however a strong physics case to discuss data preservation now, in order to allow easy access to data collected in previous years, at different centre-of-mass energy, at different pile-up conditions, or with lower trigger thresholds. Some use cases for these preservation activities can indeed become a reality in the coming year or two, requiring immediate attention. Examples of uses for these data are precision measurements with new or improved theoretical calculations, cross checks for discoveries made at higher energy/higher luminosity, studies related to new models of physics beyond the standard model. In addition to the currently ongoing and planned studies, the LHC data – being very rich in their scientific potential – will have a large physics value even after the active data-taking period.

Given the long life-time of the LHC experiments and the large volume of data collected, data preservation should be addressed during the active data-taking phase. The LHC experiments can take advantage of the experience gained during the previ-

ous experiments' data preservation activities and build on this knowledge to satisfy their own requirements. Many of the challenges are directly addressed in the experiments' computing models which are designed to distribute and store the large data volumes in the computing centres connected via grid technology.

A data preservation plan will be defined in order to prepare for the unavoidable migrations connected to software, external libraries, operating systems, storage media and the related hardware and in order to estimate the resources needed to take care of these migrations. A concrete stress-test of a plan is to consider a use-case where an analysis done on reconstructed data sets of the first years' LHC running would need to be repeated after the LHC long shut-down, foreseen during the period 2013-2014. Lessons learnt from such exercises will be incorporated in the long term preservation of the data and associated software.

As the experiments are international, with funding from agencies in Europe, the US (including both NSF and DoE) and elsewhere, this provides an excellent opportunity for international collaboration, building on the experience and challenges from previous experiments – typically with more regional funding – and extending this to a new scale in terms of data volume and duration. ILDAP would therefore address these challenges at the global level, building on these collaborations and working in conjunction with similar projects in the US and elsewhere.

While the preservation of the raw data is guaranteed by the experiments' distributed computing models (whereby multiple copies are maintained in different storage systems at various sites), the physics results are preserved through publication and stored at external, persistent repositories. This approach is already being investigated, with INSPIRE [INSPIRE] planned to be the long-term platform for such additional information.

Between the raw data and the physics results, there is much valuable knowledge and know-how worth preserving. Preserving the relevant data and information during the many intermediate steps leading from the raw data to the final physics results will require attention. Most technical facts are recorded in experiments' internal notes but many well known and well defined details such as software versions and the set of updates, conditions, corrections, the identity of events with special properties and the location of the analysis-specific code may not be explicitly recorded. As all this is known when the analysis is ongoing, it is matter of organisation and a limited amount of extra resources to preserve the full set of details. Part of the information is in collaborative media such as Twiki, posing an additional challenge to capture all relevant information. It is important that the appropriate decisions are made to define the information to be preserved and the resources for the preservation activities are made available at this early stage of the experiment's life-time. This will not guarantee that an earlier analysis can be redone in the future without technical modifications but it will guarantee that all technical knowledge connected to an analysis is preserved which is important for the internal efficiency of the experiment.

The LHC experiments will consider open access for their data with appropriate delays allowing each experiment to fully exploit the physics potential before publishing. The HEP data is complex and any public data will need to be accompanied with the software and adequate documentation. Simplified data is already provided by CMS making modest samples of selected interesting events available to educational programmes targeting high school students.

Education, training and outreach

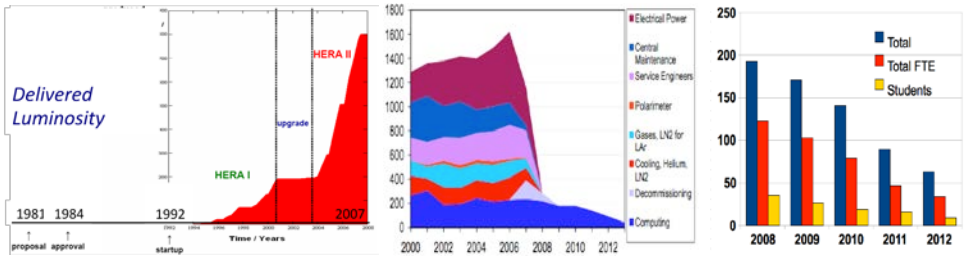
Preserving data opens new opportunities in training, education, and outreach. It permits data analysis by students whose institutes may not have originally collaborated in the experiments. This opens new opportunities for institutes in developing countries to initiate and develop HEP research. The benefit to the field is the ability to attract and train the best inquisitive minds. It also gives unprecedented opportunities to teach hands-on classes in particle physics, experimental techniques, statistics, and to explore physics topics that would not have been otherwise covered. High school students could be exposed to simplified and highly visual analyses (similar to the successful EPPOG master classes using which use special sub-sets of the DELPHI and OPAL data), in order to re-ignite the general public interest in the field and to attract new students to physics. At the same time, these data could be used by third-party applications built with the citizen-scientists in mind, following the successful example of the GalaxyZoo suite of web application [zooniverse.org]. This opportunity has been recently highlighted by the ODE project, coordinated by CERN, also funded under FP7.

1.2. ILDAP Coordination Overview

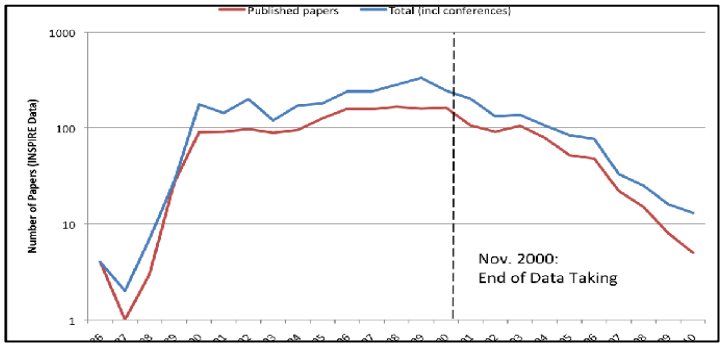
HEP is an international discipline that spans the globe. Scientists are organised into *collaborations* that are associated with (and today have the name of) a massive detector that collects data at centres such as CERN or DESY in Europe, KEK in Japan and BNL, FNAL or SLAC in the US. The largest of today's collaborations have a few thousand members and last several decades – from conception and detector design to the final analysis of the acquired data. In a sense, these communities self-organise around the idea of common shared data sets. All share the common problem outlined above. To avoid (unaffordable) duplication of effort, sharing of tools, techniques and knowledge is essential and is something that ILDAP will strongly encourage in the area of data preservation.

An analysis of the scientific potential of the preserved data can be made for experiments approaching the end of the scientific programme (for instance at HERA and b-factories). It is a fact that a dilution of the person power delays the production of some important scientific results, while some other subjects – made possible by the successive improvements of the data quality – are not addressed at all. This phenomenon has also been observed at LEP, where the publication tail exceeds ten years, with some important subjects already re-analysed. On this basis, an enhancement of the order of 5-10% is expected as a minimum if the data is preserved and the full analysis capability is maintained.

As the figures below show, accelerators typically deliver their best towards the end of the running period, whilst in contrast the funding (middle plot) and manpower (right-hand plot) decay rapidly once data taking is over (2008 in this case).



On the other hand, a plot showing the total number of published papers, in this case for the LEP collider at CERN, shows that this continues well after the end of data taking (in this case for one decade already).



The costs of preservation programmes are very small when compared to existing investments and in particular the costs of construction and operation.

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 install data preservation projects in order to avoid catastrophic loss of data once major collaborations come to an end. 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, already prefigured by the ICFA Study Group, requires a Project Manager 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. 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 and the activity of these projects will be steered by the DPHEP organisation.

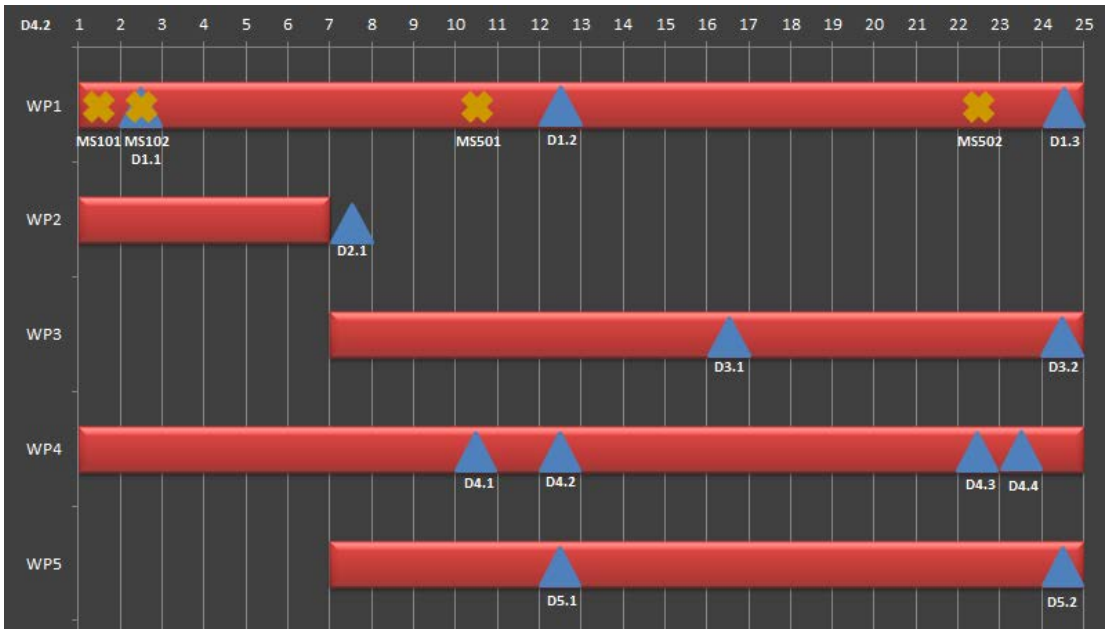
1.3. ILDAP Workplan

Introduction
Data forms a vital part of our cultural and scientific heritage that needs to be preserved for future use, including (re-)analysis. Whilst this need has been identified in numerous scientific domains and beyond, this proposal focuses on the needs of the High Energy Physics (HEP) community and in particular the recommendations of the Study Group for Data Preservation in HEP (DPHEP), which in turn also builds on the lessons learnt by the PARSE.Insight and ODE support actions, running from 2008 to 2012. Through a series of multi-disciplinary international workshops, this group identified two main use cases: educational and scientific data analysis. It calls for global sustainable data preservation in HEP. To address this need, we propose to establish a project that will interact with scientific collaborations, institutes and complementary projects in the US. The project would start by summarising and eventually extending the requirements analysis performed to date, subsequently propose standards in the needed areas, such as for data formats and for the specification of complex metadata, address concerns related to the tension between Open Data and the need of data producers to build their academic career, and finally build a range of prototypes to demonstrate the feasibility of the proposed solutions. Throughout the duration of the project, international networking both within HEP and with partners facing similar problems – possibly utilising small data volumes but for whom the period for which the data would need to be preserved can be much longer than for HEP – would form the backbone of the project, leveraging the participation of CERN, one of the partners, in the APARSEN Network of Excellence in Digital Preservation, also funded by this directorate.

Overall Strategy
The overall strategy is to build on existing work, in particular that performed by the DPHEP consortium, and take it to the next stage, establishing first a demonstrator and then a prototype of a system that supports long-term data preservation and enables re-analysis of the data. As such there are preparatory work packages (WP2 to summarise the requirements – at least in the context of the further work performed in this project; followed by WP3 to build on these requirements and define the standards that will be used for the prototyping activity). Following on from this work there is a work package devoted to prototypes (WP5), which should deliver as an interim goal a demonstrator, showing the feasibility of a long-term data preservation system and as a final goal a prototype. Throughout the duration of the project there is a networking work package (WP4). This will continue to strengthen collaboration within the HEP community as well as initiate further discussions with other communities facing similar problems and undertake the all important tasks related to dissemination of the results.

Timing

The timing of the various work packages is shown in the diagram below. Work packages 1 (project management) and 4 (networking) run for the duration of the project. Work package 2 (requirements runs for the first 6 months of the project and is followed by work packages 3 (standardisation) and 5 (prototypes).



Key

Milestone Deliverable

List of Milestones

MS101 – Project Mailing lists
MS102 – Project Templates
MS501 – Demonstrator Pre-release
MS502 – Prototype Pre-release

List of Deliverables

D1.1 – Project Website
D1.2 – Project Annual Report
D1.3 – Project Annual Report
D2.1 – Requirements Summary
D3.1 – Interim Report on Standardisation Activities
D3.2 – Final Report on Standardisation Activites
D4.1 – International Workshop on Scientific Data Preservation
D4.2 – Report Summarising Key Findings from annual workshop
D4.3 – International Workshop on Scientific Data Preservation
D4.4 – Report Summarising Key Findings from annual workshop
D5.1 – Demonstrator
D5.2 – Prototype

Table 1.3 a: Work package list

Work package No	Work package title	Type of activity	Lead partici-pant No	Lead participant short name	Person-months	Start month	End month
WP1	Project Management	MGT	1	CERN	26	1	24
WP2	Requirements	COORD	3	DESY	10	1	6
WP3	Standardisation	COORD	2	CNRS	36	7	24
WP4	Networking	COORD	1	CERN	30	1	24
WP5	Prototypes	COORD	3	DESY	54	7	24
		TOTAL			156		

Table 1.3 b: Deliverables List

Del. no.	Deliverable name	WP no.	Nature	Dissemination level	Delivery date
D1.1	Project Website	1	O	PU	M2
D1.2	Project Annual Report	1	R	PU	M12
D1.3	Project Annual Report	1	R	PU	M24
D2.1	Requirements Summary	2	R	PU	M7
D3.1	Interim Report on Standardisation Activities	3	R	PU	M16
D3.2	Final Report on Standardisation Activities	3	R	PU	M24
D4.1	International Workshop on Scientific Data Preservation	4	O	PU	M10
D4.2	Report Summarising Key Findings from annual workshop	4	R	PU	M12
D4.3	International Workshop on Scientific Data Preservation	4	O	PU	M22
D4.4	Report Summarising Key Findings from annual workshop	4	R	PU	M23
D5.1	Demonstrator	5	D	PU	M12
D5.2	Prototype	5	P	PU	M24

Table 1.3 c: List of milestones

Milestone number	Milestone name	Work package(s) involved	Expected date	Means of verification
MS101	Project Mailing lists	WP1	M1	Mailing lists established and in use
MS102	Project Templates	WP1	M2	Templates available and used for all project documents / presentations
MS501	Demonstrator Pre-release	WP5	M10	Demonstrator pre-released for evaluation
MS502	Prototype Pre-release	WP5	M22	Prototype pre-released for evaluation

Table 1.3 d: Work package description

Work package number	WP1	**Start date or starting event:**		M1			
Work package title	Project Management						
Activity Type	MGT						
Participant number	1	2	3				
Participant short name	CERN	CNRS	DESY				
Person-months per participant:	24	1	1				

Objectives

- Manage and monitor progress towards stated goals.
- Coordinate interactions with the European Commission.
- Ensure effective communication between project participants and between ILDAP and related projects.
- Provide administrative support to ensure timely, high-quality technical and financial reporting.

Description of work

This work package will be responsible for:

- All reporting to EU – the various milestones, deliverables and annual project review;
- Organising and chairing the two bodies foreseen within the project, the Project Management Board and the Technical Management Board. The Project Management Board (PMB) consists of a representative of each partner and is chaired by the Project Coordinator. Its purpose is to ensure that the project is on track with respect to its objectives and deliverables. The Technical Management Board will consist of the Project Coordinator and representatives from the work packages and will be responsible for following the progress of the project with respect to the defined work plan.

Deliverables

D1.1: Project website established. (PM2).

D1.2: Annual Report describing the progress made by the project during the first year. (PM12).

D1.3: Final Report describing the progress made by the project during the second year. (PM24).

Table 1.3 d: Work package description

Work package number	WP2	**Start date or starting event:**			M1	
Work package title	Requirements					
Activity Type	COORD					
Participant number	3	1				
Participant short name	DESY	CERN				
Person-months per participant:	6	4				

Objectives

- To understand, gather and summarise the requirements of data preservation.
- To identify the technological challenges in data preservation and explore the existing different technologies for data preservation, such as virtualisation, archival systems and validation frameworks.
- To explore the use of common simplified formats for Open Access and outreach initiatives as well as the use of meta-data across multiple scientific domains.
- To extend and enrich the variety of experiments included in existing data preservation initiatives.

Description of work

The role of this WP is to summarise and extend the requirements captured to date necessary to form a data preservation infrastructure across multiple scientific domains. This work will evaluate and build upon the findings of the Data Preservation in High Energy Physics (DPHEP) Study Group, integrating knowledge and best-practices from further scientific domains, it will also include the planning and evaluation of resources required to ensure the short and long-term availability of data, in addition to the proper archiving of the data for the longer term. The use of common and/or simplified data formats will also be examined. Given the significant role DESY plays in DPHEP, and the experience gained in this area, they will lead this WP, with additional contributions from CERN through its leading role in similar support actions (PARSE.Insight and ODE) as well as the APARSEN Network of Excellence, all funded under the same scheme.

Deliverables

A report summarising the findings of the WP on the requirements for data preservation will be prepared (D2.1), defining the concrete action to be taken and followed up with WP3. The report will be delivered in M7.

Table 1.3 d: Work package description

Work package number	WP3	**Start date or starting event:**		M7			
Work package title	Standardisation						
Activity Type	COORD						
Participant number	2	3	1				
Participant short name	CNRS	DESY	CERN				
Person-months per participant:	17	11	8				

Objectives

- Explore the use of common simplified formats for Open Access and of the different technologies for data preservation
- Identify standards on the preservation of data and documentation of critical know-how
- Propose prototypes for a standardisation procedure in HEP as related to Data Preservation activities
- Define, via interim and final report, standards to be use in prototypes

Description of work

As a natural continuation of WP2, the aim of WP3 is to identify or define standards in data preservation that will be further used in WP5 (prototypes).

This work includes the following areas:

- Data format: choice of format and migration capability to new formats;
- Data storage: data accesses and protocols;
- Data management: Metadata and know-how;
- Data access: not only access methods but also authorisation and authentication issues;
- Analysis platform: virtualisation, access interface.

This study and the necessary tests will be done in close collaboration with host laboratories and research institutes in Europe and the US. Target experiments include the LHC experiments at CERN, H1 and other experiments at DESY, BaBar at SLAC and the FNAL Tevatron experiments.

Deliverables

Reports: interim report (PM16), final report (PM24)

Table 1.3 d: Work package description

Work package number	WP4	**Start date or starting event:**			M1		
Work package title	Networking						
Activity Type	COORD						
Participant number	1	3					
Participant short name	CERN	DESY					
Person-months per participant:	18	12					

Objectives

- Harmonise and synchronise preservation projects across all stakeholders and collaborate with relevant initiatives from other fields
- Ensure the active and continuous involvement of the experiments and key host institutes in all relevant aspects of data preservation and in particular monitor and encourage progress outside of the periods immediately before and after major workshops
- Organise workshops to enlarge the community involved in the preservation activities and encourage the involvement of the different experiments
- Open the discussion to consider input from other disciplines and to share with them our knowledge and experience

Description of work

The tasks carried out by this work package will focus on coordinating the various HEP experiments, the institutes and all associated efforts on Data Preservation for Long-Term Analysis. This will cover the main goals identified by DPHEP, which are listed below.

1) Position itself as the natural forum for the entire discipline 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-wise preservation tools
3) Harmonize preservation projects across all stakeholders and liaise with relevant initiatives from other fields
4) Design the long-term organisation 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

Deliverables

Other: International Workshops, involving EU, US and others

Reports on annual workshop on Scientific Data Preservation (M12, M24)

Table 1.3 d: Work package description

Work package number	WP5	**Start date or starting event:**		M13			
Work package title	Prototypes						
Activity Type	COORD						
Participant number	3	2	1				
Participant short name	DESY	CNRS	CERN				
Person-months per participant:	18	18	18				

Objectives

- The development of a prototype validation framework, which is required to prolong the ability to perform meaningful tasks with preserved data.
- The parallel development of a data archival system suitable for long term data storage.
- To research and examine the issues associated with documentation and high level objects.
- To propose a common interface for outreach projects using preserved data and based on common standards, formats and meta-data as defined in WP3

Description of work

This WP will investigate current and future technological solutions in order to provide prototypes of systems for long term data preservation. A key project in this WP, which will build upon the work done within DPHEP, is the development of a prototype common framework to test and validate the software and data of an experiment against changes and upgrades to the environment as well as the changes to the experiment software itself. In a parallel project, the technologies required to ensure long term data integrity will also be examined and a prototype system proposed. Taking in the standards for data formats and associated meta-data, the ultimate goal of this WP is to provide a sustainable common prototype infrastructure for data preservation, data exchange and re-use across multiple scientific domains. The DESY group, which is currently in the initial development phase of such systems, will lead this project. However, given the scope and the central nature of this WP, all participants will contribute significantly.

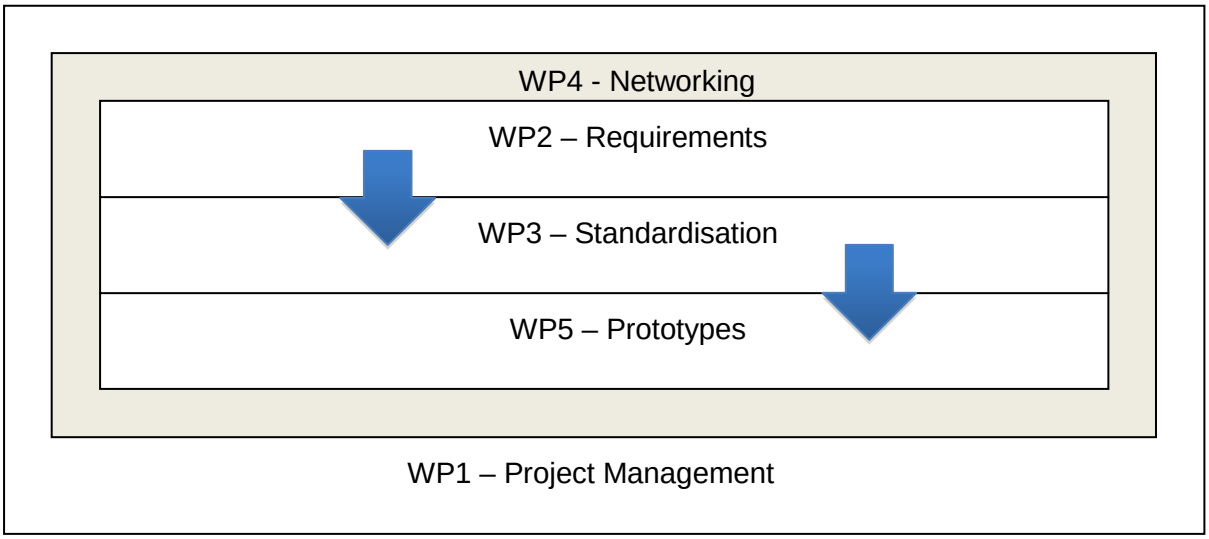
Deliverables

A Demonstrator of solutions to the technological problems will be provided by PM12, followed by a Prototype system in PM24. Pre-releases will be made available approximately 2 months prior to each workshop (see table of milestones).

Table 1.3 e: Summary of staff effort

Participant no./short name	WP1	WP2	WP3	WP4	WP5	Total person months
1 CERN	**24**	4	8	**18**	18	72
2 CNRS	1		**17**		18	36
3 DESY	1	**6**	11	12	**18**	48
Total	26	10	36	30	54	156

Figure 1 – Interdependencies between Work packages



1.3.1. Risks and Contingency Plans

The main risks and associated contingency plans are described by work package in the table below.

Table 1 – Risks and Contingency Plans

Work Package	Risk	Impact	Probability	Mitigation
WP1	Inability to manage consortium	Possible failure to meet targets, e.g. milestones and deliverables.	LOW	The consortium is small and the partners have a good track record of working together. Possible issues can always be escalated to existing, overarching and independent, bodies, where all stakeholders are represented, such as ICFA.
WP2	Lack of agreement on requirements	Project timeline maybe delayed	LOW	The DPHEP study group has already done important work in this area on which the consortium can build.
WP3	Lack of agreement on standards	Project timeline maybe delayed	LOW	Reach consensus on a limited set of standards, and provide prototypes to validate the value of standardisation.
WP4	Lack of buy-in from LHC experiments	Project scope would be limited	MEDIUM	The LHC experiments have already manifested interest (also signified by letters of support of the US partners) in the operation. The role of CERN as host laboratory of the LHC will facili-

				tate the process
WP5	Failure to deliver demonstrator	Project scope and timeline would be affected.	MEDIUM	The existing work done for the BaBar and HERA experiments are valuable proof of concept exercises in this area, and expertise can be called upon to avoid pitfalls in the starting phase of the process.

2. Implementation

2.1. Management structure and procedures

The ILDAP consortium consists of 3 partners that have a long history of working closely together on a variety of technical topics. As such, a simple management structure is considered appropriate. It is therefore proposed to establish only two boards within the project – the Project Management Board (PMB) consisting of one representative of each organisation and chaired by the Project Coordinator (PC) and a Technical Management Board (TMB), consisting of the leaders of each work package and again chaired by the PC.
The Project Coordinator will be assisted by the Coordinating Partner and an Administrative Assistant in daily administrative and financial management.
These bodies would hold regular phone or video conferences in addition to technical meetings organised within the work packages themselves. It will be a fundamental principle that all meetings will permit remote participation, given the distributed nature of the consortium and the importance of collaborating with complementary projects within the US.

2.1.1. Project coordinator

The Project Coordinator will ensure that the project meets all its contractual obligations (including all reports and deliverables), that the participants execute the defined work plans, and that the project ultimately achieves its goals. The Project Coordinator interacts with the following bodies:

- European Commission: The Project Coordinator will be the sole liaison with the European Commission for the project.
- Project and Technical Management Boards: The Project Coordinator will chair the Project (PMB) and Technical Management Board (TMB).

Dr. Jamie Shiers / CERN is proposed to be the ILDAP Project Coordinator.

2.1.2. Coordinating Partner

The Coordinating Partner is responsible for the scientific coordination, administration, and financial management of the project. The Coordinating Partner will be responsible for the distribution of the EC financial contribution to the project's partners.
CERN is the Coordinating Partner for ILDAP. Having led the European DataGrid (EDG) project as well as the EGEE series of projects, it has extensive experience in European Framework Programme projects and in performing the administrative, legal and financial services necessary to ensure the effective management of the project and the coordination of the consortium.

2.1.3. Administrative Assistant

The Administrative Assistant (WP1) will help the Project Coordinator by dealing with everyday tasks related to the management, administration, and financial reporting aspects of coordinating the project. Specifically, these tasks include arranging meetings, taking minutes, and disseminating information to the project participants and partners.

2.1.4. Project Management Board

The Project Management Board (PMB) consists of a representative of each partner and is chaired by the Project Coordinator. Its purpose is to ensure that the project is on track with respect to its objectives and deliverables and guarantee the involvement of the partners during the course of the project.

Meetings will take place at least once every quarter and may be held physically or via telephone or video conferencing. Additional meetings may be called if necessary. The agenda must be provided at least one week prior to the meeting and must include a project status report from the Project Coordinator.
All Parties shall agree to abide by all decisions of the PMB. All disputes shall be submitted in accordance with the provisions of the Grant Agreement and Consortium Agreement.

2.1.5. Technical Management Board

The Technical Management Board (TMB) will consist of the Project Coordinator and representatives from the work packages. The Project Coordinator will be the Chair of the TMB.
Meetings will be held fortnightly and may either take place physically or via telephone or video conferencing. The Chair of the TMB will prepare the agenda of the meeting in consultation with the members of the TMB. Minutes and actions from the meeting will be made available to participants before the next meeting.
The TMB is responsible for following the progress of the project with respect to the defined work plan, raising any issues (internal and external) encountered, and ensuring that other members are aware of significant events in each activity. The TMB is also responsible for the approval of deliverables and milestones.
Decisions will generally be made by consensus. Where no consensus can be reached the issue will be forwarded to the PMB for discussion and a decision.

2.2. Individual participants

2.2.1. CERN

Brief description of legal entity

CERN is an International Organisation with its headquarters in Switzerland and is the largest particle physics laboratory in the world. It is currently exploiting the Large Hadron Collider (LHC) that has just completed its second year of proton-proton running. The LHC is the world's most powerful accelerator and provides research facilities for several thousand researchers from all over the globe. The LHC experiments are designed and constructed by large international collaborations and will collect data over a period of 10-15 years. Up to 1 million computing tasks are run per day with some 15 petabytes of data generated per year. Over half a century ago the CERN charter enshrined that "… the results of its experimental and theoretical work shall be published or otherwise made generally available". Today, CERN plays a leading role in both the European and worldwide Open Access movements through a number of initiatives and FP7 projects.

Main tasks in project and relevant experience

CERN will lead the proposed project as well as work package 4 (networking). It will participate in all other work packages. Given the long lifetime of the LHC experiments and the large volume of data involved, preservation has to be addressed during the active data-taking phase. CERN has prominently contributed to a number of EGEE-related grid projects and currently leads the Services for Heavy User Communities work package of EGI-InSPIRE. Under FP6 and FP7, the IT department has been involved in some 20 European Commission-funded projects. The department has been at the forefront of computing for many years in all aspects including storage and data management.

In addition the CERN Open Access team will provide expertise from the PARSE.Insight project on drivers and barriers in data preservation, the ODE project, on opportunities offered on Open Data across disciplines and around the world, and create a link to the APARSEN Network of Excellence in digital preservation, enabling cross fertilisation of best practices.

Profiles of individuals undertaking the work

Dr Jamie Shiers leads the Experiment Support group in CERN's IT department. He has participated in several European grid projects—EGEE, EGI_DS, EGI-InSPIRE, EnviroGRIDS, PARTNER and ULICE. He has also been involved in the database, data and storage management fields, being the Vice-Chair for European Activities of the IEEE Computer Society Committee on Mass Storage Systems.

Dr Andrea Valassi leads the WLCG Persistency Framework project of the Applications Area and is a leading expert on detector-related metadata. He has concrete experience in large-scale data migrations, including those involving change of storage media, data format and of application re-implementation.

Dr Salvatore Mele is head of Open Access at CERN. He is the interim project manager for the emerging SCOAP3 consortium, aiming to convert all HEP literature to Open Access. As strategic director of INSPIRE, the Invenio-based digital library for HEP, he is exploring links between digital libraries, e-Infrastructure, data preservation and sharing. Dr. Mele is also the coordinator of the SOAP and ODE FP7 projects, and was one of the architects of the OpenAIRE and OpenAIREPlus initiatives. He holds a PhD in Physics and previously worked at the CERN LEP accelerator, where he led teams that measured fundamental physics constants, hunted for the Higgs boson and searched for hints of extra dimensions.

In addition, CERN plans to hire on project funds the necessary manpower to cover project administration and financial reporting as well as two staff with experience in data preservation for its involvement in the other work packages.

2.2.2. CNRS

Brief description of legal entity

The Centre National de la Recherche Scientifique is a government-funded research organisation, under the administrative authority of France's Ministry of Research. Its annual budget represents a quarter of French public spending on civilian research. As the largest fundamental research organisation in Europe, CNRS carries out research in all fields of knowledge, via its eight CNRS Institutes: Institute of Chemistry (INC), Institute of Ecology and Environment (INEE), Institute of Physics (INP), Institute of Biological Sciences (INSB), Institute for Humanities and Social Sciences (INSHS), Institute for Computer Sciences (INS2I), Institute for Engineering and Systems Sciences (INSIS), Institute for Mathematical Sciences (INSMI) and its two national institutes with national missions, the National Institute of Earth Sciences and Astronomy (INSU) and the National Institute of Nuclear and Particle Physics (IN2P3). CNRS maintains its own laboratories as well as joint laboratories with universities, other research organisations and industry throughout France and overseas in several countries. Measured by the amount of human and material resources it commits to scientific research or by the great range of disciplines in which its scientists carry on their work, the CNRS is clearly the hub of research activity in France. It is also an important breeding ground for scientific and technological innovation, and has been one of the most active participants to previous and current European Framework Programme, including in a coordinating role.

Main tasks in the project and relevant experience

The Laboratory involved in the ILDAP project is CC-IN2P3, the National Computing Centre of IN2P3. It is a CNRS Service and Research Unit with a staff of 85 people, including 57 high level computing engineers. CC-IN2P3 operates the largest computing centre in France dedicated to data processing. It provides computing and storage services for scientific research needs, mainly in the field of HEP. More than 2000 users working on 40 international experiments use its services. As a Tier1 for the LHC experiments, it has contributed to many of the major EGEE and WLCG projects. It plays a central role in the operation of the GRID at national (France-Grilles) and international (European project EGI-InSPIRE) levels.
For many years, CC-IN2P3 has opened its computing capabilities and expertise to multidisciplinary activities (biomedical, biology, physics, humanities, etc.).
CNRS will lead WP3 on standardisation and also participate in WP6 on prototypes.

Profiles of individuals undertaking the work

Dr. Ghita Rahal currently leads the Support Group of CC-IN2P3. She graduated in the field of experimental HEP and participated as a physicist in various experiments at CERN and FNAL. Besides her research activity, she has led various teams in charge of detector maintenance and calibration, software and physics analysis and has been part of the board of referees for publications. She is currently a member of the ATLAS collaboration at CERN.
Dr. Cristinel Diaconu is the spokesperson of the H1 experiment at the HERA facility at DESY and a member of the Centre de Physique des Particules de Marseille (CPPM), where he acts as "President du Conseil Scientifique du CPPM". He has played a leading role in the DPHEP organisation, including the development of the interim report and the "Blueprint document" that is currently being finalised. He is "Charge de mission pour l'informatique" at IN2P3 and supervised IT services over a network of 20 laboratories and 280 engineers. He is also the Chair of the Evaluation and Survey Committee of the CC-IN2P3 and Chair of the Oversight Board of the GEANT4 project.
Additional staff: CNRS will also hire additional staff on project funds as per the work package breakdown tables in section 1.

2.2.3. DESY

Brief description of the legal entity

DESY is one of the world's leading centres for the investigation of the structure of matter. DESY develops, operates, and uses accelerators and detectors for photon science and particle physics. As a member of the Helmholtz Association in Germany, DESY is a non-profit research organisation supported by public funds. With its more than 50-year success story in particle research and its unique facilities, DESY has played a decisive role in particle and astroparticle physics. International cooperation across cultural and political boundaries enjoys a long tradition at DESY. DESY participates in a number of international facilities which are no longer supported by one country alone, but are instead realized as wide-ranging international projects.
Now, DESY is involved in the experiments at the world's most powerful accelerator, the Large Hadron Collider (LHC) in Geneva, Switzerland. In addition, computing centres for monitoring LHC data acquisition and data analysis are being established at DESY. DESY is also playing a leading role in particle physics' large future project, the planned International Linear Collider ILC. The ILC is based on the superconducting accelerator technology developed and tested by DESY and its international partners.

Main tasks in the project and relevant experience

DESY is the host laboratory of several HEP experiments, including those exploiting the data from the HERA accelerator. DESY also provides a large computing facility used extensively by the HEP experiments at the LHC. Today the DESY Tier-2 GRID infrastructure constitutes the largest resource of its type, which is supplemented for analysis by the National Analysis Facility, also hosted at DESY.
The IT division is actively collaborating with the HERA experiments within the framework of the on-going data preservation activities at DESY. Having taken a leading role in the global DPHEP initiative since it began in 2009 DESY, continues to support this international effort, with joint projects underway between all groups, under the guidelines of the recommendations of DPHEP.
DESY will also be the host laboratory for future accelerator programmes in photon science, with the need to store and preserve data volumes larger than or comparable to the amount of data expected from the LHC.

Profiles of individuals undertaking the work

Dmitry Ozerov has been working in HEP as a research scientist for more than a decade. Before joining the DESY-IT division in 2008 he lead the software development project for one of the largest active experiments in HEP at that time. He actively participates in EGEE-III and EGI projects, ensuring the sustainability of the DESY analysis infrastructure for the HEP community.

Dr. David South is a member of the H1 and ATLAS collaborations, graduating in experimental HEP in 2003. Since joining H1 in 2000 he has been involved in the development of the analysis software environment and is computing coordinator of the experiment since 2008. Heavily involved in the DPHEP initiative since its conception, he now leads the data preservation group at DESY.

Additional staff: DESY will hire two staff on project funds as per the work package breakdown in section 1.

2.3. Consortium as a whole

The ILDAP consortium consists of 3 funded partners; 2 in EU Member States (FR, DE) and the 3rd in an Associated Country (CH). All parties have been involved in the Data Preservation in HEP activity since its onset and, as either accelerator laboratories (CERN and DESY – where the data is produced) and/or users of the facilities (CNRS) have strong motivation to find a long-term solution to the problems associated with data preservation for future analyses. Both CERN and DESY have considerable experience in the data management and storage domain as software providers (CASTOR, DPM and EOS from CERN, dCache from DESY and the dCache consortium). Collectively, these storage solutions manage much of the LHC data that currently exceeds 100PB[2] worldwide. CNRS, through its WLCG Tier1 site at IN2P3, is also a world-class service provider in this domain, serving all 4 LHC experiments plus numerous others, including those that took data at US facilities, such as BNL, FNAL and SLAC. CNRS, through its site at LAL, has also contributed to the development and testing of DPM, the storage solution primarily used at Tier2s. In other words, the project partners have established relationships and a proven track record of working effectively together on common goals.

CERN is proposed as Coordinating partner and has extensive experience in this respect, having been involved in the complete series of EU-funded grid infrastructure projects (EDG and EGEE I-III as lead partner), as well as numerous other EU-funded projects. CERN also leads the Worldwide LHC Computing Grid, which includes partners worldwide, and coordinates the associated service. As the host laboratory for the LHC experiments, CERN will have to preserve the associated data for at least the duration of data taking of this machine, expected to exceed two decades. It has already had to face a number of the problems related to data preservation, not only for the current experiments, but also those at the previous collider, LEP, for which reanalysis is currently on-going, motivated by the apparent low mass of the Higgs boson based on analyses from the LHC as well as FNAL's Tevatron collider. CERN will lead the work package on networking (WP4), leveraging on its unique network which reaches virtually every single HEP scientist of the planet. It will participate in all other work packages, in particular WP5 on prototypes with a special focus on the needs of the LHC experiments. CERN will contribute its unique expertise on Open Access to advise on policy and data-governance matters in the Requirements work package (WP2) and provide cross-fertilising links from the PARSE.Insight project on drivers and barriers in data preservation, the ODE project, on opportunities offered on Open Data across disciplines and around the world, and create a link to the APARSEN Network of Excellence in digital preservation.

CNRS is proposed to lead the project in the Standardisation work package (WP3). CC-IN2P3 is already bringing solutions to data storage and analysis for various types of experiments and fields of research. This will include exploring the suitability of the different technologies involved in data preservation and the use of the various common simplified formats for the needs of the different experiments. By identifying standards in this area, the working group will establish the corresponding future working directions. CNRS will also contribute to the development of technological solutions in the Prototypes group (WP5).

DESY is proposed to lead the project in two areas: Requirements (WP2) and Prototypes (WP5), as well as participating in Standardisation (WP3) and Networking (WP4). Over the last few years DESY has participated in the DPHEP organisation,

[2] Other solutions include StoRM, BeStMan and xrootd.

establishing the schema and gathering the knowledge required to produce early results in the development of data preservation projects. The requirements and standards of preservation of HEP data have been defined in no small part by surveying the large and varied quantity of HEP data involved at DESY, as well as an appreciation of the technological models involved. The DESY group is already in the process of producing real solutions to the problems related to data preservation. After a successful initial pilot phase, the group at DESY now implement a full-scale project to not only ensure the integrity of the HERA data at DESY but also to validate the analysis software against future changes. The prototype model already in development is by design extendable to other HEP data.

2.4. Resources to be committed

The EU funded resources will be assigned to the partners based on the Work Package breakdown outlined in Section 1.

The US partners include the Open Science Grid (OSG), the main laboratories in the US (BNL, FNAL and SLAC) and their associated experiments (that themselves have both US and European partners) as well as the US components of the ATLAS and CMS experiments, centred around BNL and FNAL respectively. Through OSG and its stakeholders, a project proposal is being prepared for submission to the NSF in early 2012 to work directly with ILDAP. The two projects will work together with DPHEP and partners in Europe, the US and Asia-Pacific.

With the exception of the administrative and financial tasks (24 person months), assigned to the coordinating partner, the work shall be technical and performed in close collaboration with the HEP experiments and the associated laboratories and institutes worldwide (132 person months). As a coordination activity, it relies on additional activities that are external to the project – in particular work performed within each experiment related to Data Preservation (estimated by the DPHEP study group to be of the order of 2-3 FTE per experiment) as well as in collaborating and complementary projects in the US, funded by the National Science Foundation (NSF) or Department of Energy (DoE).

In addition, the role of Project Coordinator would be funded by CERN and not via EU funds. However, it is important to stress that this additional effort is considered fundamental to carry the valuable work of this study group to the next stage and to ensure full cooperation with the US in this important area.

The work of this project would be to coordinate, facilitate and enable this experiment-specific work through the work packages that are defined, such as requirements gathering, definition of the relevant standards and the development of appropriate prototypes. The hardware resources for the necessary prototypes are modest in size compared to those needed to support the ongoing data taking, processing and analysis activities of the laboratories involved in this project and will be absorbed by the partners concerned.

For the output of the project to have any value, it must be accepted by the global community, which includes the sites and collaborations (experiments) involved. The work of the project would be regularly presented to the communities using existing meetings, workshops and conferences as well as at dedicated annual and topical workshops within the context of the project itself. Suitable external events include the HEPiX forum and the Computing in High Energy Physics (CHEP) conference series. The budget calculations set aside the necessary resources for the organisation of the workshops foreseen in the list of deliverables (M 10 and 22), as well as to allow the participation of the hired staff in the mentioned conferences.

3. Impact

3.1. Expected impacts listed in the work programme

The need for data preservation in HEP has been discussed extensively in section 1 of this proposal. Here we will focus on the potential impact through the development, dissemination and use of the results of the ILDAP project.

The major impact of the project will be to allow the possibility of long-term completion and extension of scientific programmes in the HEP scientific community. Natural continuation of the programmes of the different organisations will ensure the full exploitation of the potential of the data at a time when the collaboration has diminished or even dissolved. The idea of performing cross-collaboration analyses is also a very important benefit that will arise from this project, where the comprehensive and coherent analysis of several experimental data sets opens up appealing scientific opportunities to reduce the uncertainties of single experiments, or provide the means to do ground-breaking combinations of experimental results. Finally, several scientific opportunities are available by re-using data from past experiments. New theoretical developments can be probed with the data of an experiment that is no longer running and whose data are from a kinematic region not accessible at present day facilities.

Several stakeholders clearly emerge as participants in data preservation activities in HEP, such as the scientific collaborations, the host laboratories – such as CERN and DESY in Europe, the computing centres – such as CCIN2P3/CNRS, and the national funding agencies. All these bodies have invested a vast amount of resources to achieve a wealth of scientific results, and begin now to invest additional resources into finding solutions for data preservation. There are a number of different initiatives to preserve this data and ensure the capability of its future analysis, mainly taking place within the DPHEP study group. However, within this group there are not enough resources and the contributors are isolated from each other at an experiment or even laboratory level, lacking coordination among the initiatives.

The project starts with two technical work packages running in parallel, the Networking WP (WP4) and the Requirements WP (WP2), described below.

We believe that the coordination to be done by the ILDAP project, within the **Networking WP**, will allow a collective focus, effective transfer of knowledge and the development of scalable solutions among the different stakeholders, avoiding duplication of effort across different initiatives and encouraging a collective approach to this challenge. The ILDAP coordinated effort will allow concrete results to be achieved and a holistic long-term sustainable plan, keeping limited financial resources allocated to the data preservation activity. The partners participating in this proposal have strong connections with the CERN LHC experiments as well as other High Energy Physics programmes in Europe and the United States. At the same time they are heavily involved in supporting non-HEP disciplines in the scope of the EGI-InSPIRE programme. This will put them in a unique position for the coordination of a cross discipline data preservation effort and directly address the concern above – namely the missing effort and lack of coordination across initiatives.

This initiative aims at building on the global knowledge established within DPHEP from the existing High Energy Physics preservation efforts in the EU and United States. This is the role of the **Requirements WP**, which will expand the scope of data preservation beyond HEP, by applying and adapting the DPHEP conclusions to other fields.

In a second phase, the **Standardisation WP** will have an impact on the coordination of the existing initiatives aiming towards the creation of a standard, coherent with the programmes already in place within individual experiments. The experience we will gather with existing initiatives and the effort in standardising the process will allow us to deliver a sustainable system for long-term data preservation for many scientific domains, again extending the model from High Energy Physics to other disciplines.

Another relevant impact will be on the possible opportunities of **Open Access** of the preserved scientific data, to balance with the scientific practices of the community and the potential for citizen scientists to peruse HEP data. Together with important scientific drivers such as physics supervision and authorship, these issues will be addressed in the **Requirements work package**. The publication of physics results during the lifetime of collaboration follows rigorous procedures, exercised over many years. Certification mechanisms ensuring the correctness of the produced results will be therefore implemented, reflecting the quality requirements specific to the level of detail used in the analysis. The authorship procedures are also affected. Author lists of HEP publications are defined according to internal mechanisms and include usually all members of the collaboration. Beyond the lifetime of a collaboration, that typically lasts several years beyond the end of data taking, the authorship rules for use in scientific papers will be clearly defined such that data analysis is encouraged and that proper credits are allocated to the collaboration that collected the data. Those aspects will imply a new cultural approach for the scientific HEP community toward their private analyses and the documentation needed on the adopted methods and tools to guarantee that the results can be reproduced later in time.

Information management and storage will also profit from the ILDAP project, concerning the extension of public documentation, the enhancement of information by storing figures, analysis data, notes and internal legacy material. An important impact of the **Networking WP** concerns tools used by scientists to access papers and other public resources, like INSPIRE. Currently the INSPIRE repository already provides full access to indexed pre-prints and articles, as well as figures, captions and data tables extracted from these papers. Additionally, INSPIRE could also provide access to high-level data that were used to produce the results. This will be possible as a result of the standardisation of the high-level data format, as defined in the **Standardisation WP** of the ILDAP project. The ability to accurately attribute sources and distribute scientific credit is a potential benefit of data preservation. They constitute a clear added value for funding agencies that are increasingly paying attention to additional methods of impact assessment.

The **Prototyping WP** will affect the development/use of new technologies for long-term data storage and analysis software preservation. The main aspects will be related to virtualisation techniques and virtual repositories, data and analysis migration procedures, data validation suites and archival infrastructures. Solutions for automatic migration to new media generations and technologies, as well as for automatic data integrity checks will be designed that will also be of benefit to other scientific communities.

As an example of the potential impact of this project we can quote the recent resurrection and re-analysis of data from JADE, an experiment that operated at the PETRA electron-positron collider between 1979 and 1986. Applying new theoretical input and new experimental insights and methods, the old data provided new physics results in an energy range which today is not otherwise accessible and also allowed combined analyses with data from more recent experiments. This re-analysis of data that is more than twenty years old was made possible by the commitment of a few individuals. It has been a tour de force and far from a standard enterprise in HEP.

The analysis of this example shows that the preservation of HEP data at the highest level can be successful in the presence of proper means. The definition of a standard for data preservation would allow this kind of analysis to be done at considerably lower cost.

Through the DPHEP organisation and via collaboration with complementary US projects, such as the foreseen Data and Software Preservation for Open Science (DASPOS), as well as through the international HEP experiments and through the various discipline-oriented and technical conferences (such as IEEE Nuclear Science / Mass Storage Symposia), the project will both inform and take account of external related research activities. Although the primary focus of ILDAP is on the needs of the HEP community, it will also retain links with other disciplines, such as Astronomy / Astrophysics as well as Life and Earth Sciences and in the Arts and Humanities area. The speed at which any concrete work might take off in these fields depends to a large extent on the funding that can be found within these communities.

The primary external factor that will determine whether these impacts are achieved is the necessary buy-in from the experiments and the allocation of the needed manpower within them to work on data preservation aspects (see risk table in section 1). The fact that there is support for this proposal at a high level within the experiments and the institutes, all of whom have a long track record of working together successfully, suggests that this is low risk and is in any case one of the main areas to be addressed by the networking work package, WP4.

3.2. Spreading excellence, exploiting results, disseminating knowledge

3.2.1. Project Internal

Since the final aim of this project is to bring together various collaborations, spread knowledge and define common solutions, it will be important to define a series of events where the community can meet, discuss ideas and follow the status of the activities.
The schedule of those events will naturally follow the milestones defined in the work packages:

- A kick-off meeting at the beginning of the project to assess the state of the art. This is the occasion for several communities to present what they have implemented in terms of Data Preservation and where we collect requirements for standardisation in order to define the working groups. We plan to involve all partners, including those without Data Preservation programmes, which will benefit from such projects in the future.
- Month 12 – Mid-term workshop. The Standardisation work package should report on the progress they have made so far. The work package on Prototypes should be defined based on the knowledge acquired until this point.
- Month 24 – Final workshop with the wrap up of the activities carried out during the project and with a discussion on the possible activities to be held in the future.

Given the geographical distribution of the involved partners, it will be of great benefit to set up a website to gather and collect all the relevant information. E-groups and e-fora will be defined for general discussion on the project and for each specific work package.

3.2.2. External Dissemination and Outreach

In addition to the activities within the project itself, every opportunity will be used to disseminate the work to wider audiences. This will include the IEEE Massive Storage and Technologies conference series (IEEE MSST), the IEEE Nuclear Science Symposia (held jointly with the IEEE Medical Imaging Conference – hence opening the door to Life Science activities), the Computing in High Energy Physics (CHEP) conference series and the HEPiX working group. As the institutes involved are also involved in grid and cloud-computing related activities, and as this work is relevant for users of Distributed Computing Infrastructures, it will also be disseminated through relevant channels such as EGI events in Europe and OSG ones in the US.

Moreover, in order to address the wider public, articles will be published in magazines such as International Science Grid This Week (iSGTW) and dedicated outreach events will be organised in the CERN Globe of Innovation, as well as similar events held at other European sites (CNRS, DESY and through parallel projects also in the US).

3.2.3. Educational Outreach

The project will also have a large impact in terms of educational outreach.

As our knowledge of the universe expands and new data are collected, we find it useful to return not only to our past conclusions but also to the old data themselves and check whether or not it all survives in a consistent interpretation. Having access to data from experiments all over the world can raise outreach efforts to the public to another level by letting non-experts interact with the scientific experience in a way not previously possible. The outreach tools developed for these efforts can also be used for undergraduate college courses and to train graduate students who will be the next generation of physicists at the frontier.

By improving our educational tools for the general public, we will also develop better techniques for teaching new graduate students and collaborators which will allow them to contribute more quickly to the experiments. (For example, by repeating key analyses performed in the past – the "historical base" that they need if they are expected to make the next big discoveries).

It may be that non-experts are able to provide an outside perspective which benefits the HEP community in data visualisation, algorithm development or even the scientific analysis itself.

Well calibrated, well understood datasets can address the aforementioned outreach and education efforts. There are four main groups who can learn from and benefit from these data:

1. The general public;
2. High school science students;
3. College students studying courses in particle physics or computing;
4. Graduate students in particle physics.

Each group brings its own challenges and may interact with the data in different ways.

Thus this work will raise the overall awareness and appreciation of scientific work so that it becomes an integral part of future educational models, addressing the needs identified above.

3.2.4. Policy Makers and Future Strategies / Work

It is expected that data preservation for re-use will be a cornerstone of Europe's Digital Agenda. For example, if it is required to retain digital health records for the lifetime of an individual (and potentially much longer to observe trends and other collective events), data will have to be preserved for at least several decades. Although there are specific requirements for such data that are not as relevant to HEP, handling the vast volume of HEP data will no doubt be valuable experience in preparing for long-term data preservation in other fields. For example, the total data volume in Astronomy and Astrophysics doubles every year whereas that in HEP grows linearly and is currently dominated by that generated by the LHC. Similarly, data in the Life Science domain can be expect to grow significantly as more patient information is made available digitally, such as medical imaging data. Thus, the work supported by this project can be seen as valuable input into preparing longer term and more diverse preservation programmes and in helping to identify requirements and even standards that might be applicable to multiple application domains.

4. Ethical Issues

As shown in the table below, none of the ethical issues raised are relevant for HEP data and hence not for this project.

Table 2: Ethical Issues Table

	YES	PAGE
Informed Consent		
• Does the proposal involve children?		
• Does the proposal involve patients or persons not able to give consent?		
• Does the proposal involve adult healthy volunteers?		
• Does the proposal involve Human Genetic Material?		
• Does the proposal involve Human biological samples?		
• Does the proposal involve Human data collection?		
Research on Human embryo/foetus		
• Does the proposal involve Human Embryos?		
• Does the proposal involve Human Foetal Tissue / Cells?		
• Does the proposal involve Human Embryonic Stem Cells?		
Privacy		
• Does the proposal involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)		
• Does the proposal involve tracking the location or observation of people?		
Research on Animals		
• Does the proposal involve research on animals?		
• Are those animals transgenic small laboratory animals?		
• Are those animals transgenic farm animals?		
• Are those animals cloned farm animals?		
• Are those animals non-human primates?		
Research Involving Developing Countries		
• Use of local resources (genetic, animal, plant etc)		
• Impact on local community		
Dual Use		
• Research having direct military application		
• Research having the potential for terrorist abuse		
ICT Implants		
• Does the proposal involve clinical trials of ICT implants?		
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	**YES**	

5. Annex: Details and Examples

Cross-collaboration analyses

The comprehensive analysis of data from several experiments at once opens appealing scientific opportunities to either reduce statistical and/or systematic uncertainties of single experiments, or to permit entirely new analyses that would be otherwise impossible. Indeed, ground-breaking combinations of experimental results have been performed at LEP, HERA and the Tevatron, during the collaborations' lifetime, providing new insight in precision measurements of fundamental quantities, and extending the ranges for searches of new physics. Preserved data sets may further enhance the physics potential of experimental programmes, by offering the possibility of combinations which would not be otherwise possible. Data from facilities where no active collaboration is operating would be available for combination with new data. At the same time, well-documented preserved data would also enhance opportunities for combinations among current experiments, which may be otherwise prevented by the lack of standards leading to insurmountable technical or scientific problems. The HEP community comprises sub-communities of experts in various fields such as flavour physics, neutrino physics, and so on. As clearly found by the PARSE.Insight study on opportunities in data preservation, a precursor to this project and DPHEP also funded by the same scheme, these expert communities would greatly benefit from having simultaneous access to data sets from relevant experiments. For example, B-physics experts could devise analyses simultaneously using data from BaBar, Belle and Cleo-C. Such an effort to combine analyses is already ongoing, for example between the H1 and ZEUS collaborations, and an evaluation of such an approach is underway between the Belle and BaBar collaborations. An effort in standardising and/or documenting data sets for long-term preservation would have an immediate return in facilitating these combinations.

Data re-use

While data re-use is the norm in several fields of science, from bio-informatics to Earth observation, it has historically been hampered in HEP by the lack of a concerted approach in data preservation. Several scientific opportunities could be seized by re-using data from past experiments. For instance, new theoretical developments could allow new analyses leading to a significant increase in precision for the determination of physical observables. Theoretical progress can also lead to new predictions (e.g. of new physics effects) that were not probed when an experiment was running and is not accessible at present-day facilities. Similarly, new experimental insights (e.g. breakthroughs in Monte Carlo simulation of detector response) or new analysis techniques (e.g. multivariate analysis tools, greater computing capabilities) could allow improved analyses of preserved data, with a potential well beyond the one of the published analyses. Results at future experimental facilities may require a re-analysis of preserved data (e.g. because of inconsistent determinations of physical observables, or observation of new phenomena which may/should have been observed before). For example results from the LHC experiments may very well induce re-analysis of LEP, Tevatron or HERA data.

Real Examples of data re-analysis

In spite of the fact that data preservation has not been planned in most of the experiments, examples of the usefulness of long-term access to the data and to the analysis frameworks exist and illustrate the generic research case presented above.

The reanalysis of the JADE data is a well known example of a resurrection of an almost lost data set. Advanced theoretical knowledge and analysis methods compared

to those being available at PETRA times, in particular for the modelling of the hadronic final states, lead to an improved measurement in a unique energy domain, not available and not reproducible anymore in spite of the higher energy and luminosity available at LEP. Enhanced and more profound theoretical knowledge, more sophisticated Monte Carlo (MC) and hadronisation models, improved and optimised experimental observables and methods, and a much deeper understanding and precise knowledge of the Standard Model of electroweak and strong interactions make it mandatory and beneficial to reanalyze old data and to significantly improve their scientific impact. [BETHKE]

Searches for new physics can also benefit from the re-analysis of the preserved data sets. As explained above, new models or better understanding of the theoretical framework may reveal islands of sensitivity that were not explored before. This is the reason for the recent re-analysis of the ALEPH data to search for a low mass Higgs supersymmetric partner which may be produced in pairs and would be able to decay in four tau leptons. This configuration and the corresponding decay channel were not explored during the collaboration lifetime and were shown to cover a new domain in the parameter space, i.e. a real discovery chance was explored at about ten years after the data taking period. The re-analysis involved a recovery of the analysis software and a dedicated effort to reprocess samples of Monte Carlo events, illustrating the need for preservation of the capabilities to perform complete analyses.

Another example covers the recent rise in interest in models involving the so-called dark photons. These bosons would result from a special theory extending quantum electrodynamics and leading to a heavy photon weakly interacting boson coupling to the photon. This configuration would lead to a change in the branching ratio of neutral pions to photons. These branching ratios are best measured in so-called beam-dump experiments, performed essentially at previous fixed target facilities. The re-analysis of some of these data led to improved restrictions on such models, which are nowadays theoretically allowed. It is striking to note that most of the recent exclusion analyses performed around dark photons models use experimental data that is older than two decades and in fact struggle to recover some of the analysis features (like the acceptance modeling) which are not available directly as a consequence of the data and software lost. [BLEUMLEIN]

6. Annex: References

Table 3 – References

DPHEP	ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics – http://www.dphep.org.
BETHKE	Data Preservation in High Energy Physics - why, how and when?, Siegfried Bethke (Munich, Max Planck Institute), Sep 2010. 5 pp. Published in Nucl.Phys.Proc.Suppl. 207-208 (2010) 156-159, Presented at SPIRES Conference C10/06/28.2, e-Print: arXiv:1009.3763 [hep-ex]
BLEUMLEIN	New Exclusion Limits for Dark Gauge Forces from Beam-Dump Data, Johannes Blumlein (DESY), Jurgen Brunner (DESY and CPPM), DESY-11-062, DO-TH-11-11, SFB-CPP-11-18, LPN-11-17. Apr 2011. 9 pp. Published in Phys.Lett. B701 (2011) 155-159 e-Print: arXiv:1104.2747 [hep-ex]
INSPIRE	Production service at http://Inspirehep.net
PARSE.Insight	Permanent Access to the Records of Science in Europe http://www.parse-insight.eu/
ODE	Opportunity for Data Exchange http://ode-project.eu
APARSEN	Alliance for Permanent Access to the Records of Science Network http://aparsen.eu

7. Annex: Glossary

ACE	Adaptive Communication Environment
ADAMO	Entity-relationship model
AFS	Andrew File System
AGS	Alternating Gradient Synchrotron
ALEPH	Apparatus for LEP Physics at CERN
ALICE	A Large Ion Collider Experiment; an LHC experiment
Amazon EC2	Amazon Elastic Compute Cloud
APARSEN	Alliance for Permanent Access to the Records of Science in Europe
AR	Annual Report
arXiv	arXiv is an e-print service in the fields of physics, mathematics, non-linear science, computer science, quantitative biology, quantitative finance and statistics
ATLAS	A Toroidal LHC Apparatus; an LHC experiment
BaBar	The BaBar acronym, which is the name of the experiment and detector collaboration, refers to the B/B-bar system of mesons which are produced at SLAC's PEP-II collider.
BAD	BaBar Analysis Documents
BAIS	BaBar Analysis Information System
BASF	Belle Analysis Framework
Belle	Particle physics experiment conducted by the Belle Collaboration at the High Energy Accelerator Research Organisation (KEK) in Japan.
BEPC	Beijing Electron Positron Collider
BES	Beijing Spectrometer
B-factory	A collider-based scientific machine designed to produce a large number of B mesons and analyse their properties
BNL	Brookhaven National Laboratory
BOSS	BESIII Offline Software System
B-physics	Physics based on the analysis of B mesons (see also B-factory)
CASTOR	CERN Advanced Storage Manager
CCIN2P3	Centre de Calcul de l'IN2P3
CDF	Collider Detector at Fermilab
CDST	Compressed Data Storage Tape
CEBAF	Continuous Electron Beam Accelerator Facility
CentOS	A Linux Operating System distribution based on RHEL
CERN	European Organisation for Nuclear Research
CERNLIB	The CERN Program Library is a collection of FORTRAN77 libraries and modules, currently maintained "as is" by CERN
CERN-VM	CERN Virtual Machine is a baseline Virtual Software Appliance for the participants of CERN LHC experiments
CESR	Cornell Electron Storage Ring
CHEP	Computing in High Energy Physics conference series
Cleo	General purpose particle detector at the Cornell Electron Storage

	Ring (CESR)
CLHEP	Class Library for High Energy Physics
CMS	Compact Muon Solenoid; an LHC experiment
CMT	A software configuration tool
CNAF	Centro Nazionale dell'INFN
CNRS	Centre National de la Recherche Scientifique
ConditionDB	Non-event data for monitoring the detector operation and needed for event reconstruction
ConsBlock	A 30-minute block of data (reconstruction block) produced by the BaBar online event processing software
COORD	Coordination activities
CORBA	Common Object Request Broker Architecture
CPEP	Contemporary Physics Education Project
CPU	Central Processing Unit
CREATIS	Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé
CSA-CA	Coordination and Support Actions - coordinating actions
DAQ	Data Acquisition
dCache	A mass storage solution
DDL	Data Definition Language
Deliverable Nature	R = Report, P = Prototype, D = Demonstrator, O = Other
DELPHI	Detector with Lepton, Photon and Hadron Identification
DESY	Deutsches Elektronen-Synchrotron
DIS	Deep inelastic scattering
DØ	DØ was one of two major experiments located at the the Tevatron Collider, at the Fermilab in Batavia, Illinois, USA
DoE	Department of Energy
DOI	Digital Object Identifier System
DPHEP	Study Group for Data Preservation in High Energy Physics
DPM	Disk Pool Manager
DQ	Data Quality
DST	Data Summary Table
EC	European Commission
EDG	European DataGrid
EGEE	Enabling Grids for E-sciencE
EGI	European Grid Infrastructure
EGI_DS	EGI Design Study
EGI-InSPIRE	EGI Integrated Sustainable Pan-European Infrastructure for Researchers in Europe
EnviroGRIDS	Building Capacity for a Black Sea Catchment Observation and Assessment System supporting Sustainable Development
EOS	EOS is a disk pool prototype that is under consideration for analysis-style data access
EPPOG	The European Particle Physics Outreach Group
EU	European Union
European XFEL	ESFRI project: X-ray Free Electron Laser research infrastructure

FASTJET	Software package for jet finding in proton-proton and electron-positron collisions
FERMILAB	Fermi National Accelerator Laboratory
FNAL	Fermi National Accelerator Laboratory
FP	Framework Programme
FTE	Full-Time Equivalent
GAF	General ADAMO File
GalaxyZoo	Interactive project that allows the user to participate in a large-scale project of galaxy research: classifying millions of galaxies found in the Sloan Digital Sky Survey
GDML	Geometry DescripTion Markup Language
GEANT	Detector Description and Simulation Tool
GeV	Gigaelectron Volt
GKS	Graphical Kernel System is the first ISO standard for low-level computer graphics
GLAST	Fermi Gamma-ray Space Telescope, formerly the Gamma-ray Large Area Space Telescope
GPD	Generalized Parton Densities
GridKa	Grid Computing Centre Karlsruhe
H1	Particle detector on the HERA particle accelerator at DESY
HAT	H1 Analysis Tag
HEP	High Energy Physics
HEPAP	High Energy Physics Advisory Panel
HEPData	The HEPData Project has for more than 25 years compiled the Reactions Database containing what can be loosly described as cross sections from HEP scattering experiments
HEPiX	High Energy Physics Unix Information Exchange forum
HEPSPEC	The High Energy Physics (HEP) SPEC benchmark is a set of test applications which stress the processor with operations and algorithms used commonly in applications from the physics community
HERA	Hadron-Elektron-Ring-Anlage; a particle accelerator at DESY
HERMES	Experiment investigating the quark-gluon structure of matter at DESY
HPSS	High Performance Storage Systems
HSM	Hierarchical Storage Management
HTML	HyperText Markup Language
I/O	Input/Output
ICFA	International Committee for Future Accelerators
ICT	Information and Communication Technology
IDG	Institut des Grilles
IEEE	Institute of Electrical and Electronics Engineers
IHEP	Institute of High Energy Physics
ILC	International Linear Collider
ILDAP	International Long-term Data and Analysis Preservation
IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
INFN	Istituto Nazionale di Fisica Nucleare

INSPIRE	Next-generation High Energy Physics (HEP) information system, INSPIRE, which empowers scientists with innovative tools for successful research at the dawn of an era of new discoveries.
INSU	Institut national des sciences de l'Univers
IP	Intellectual Property
IR2	Interaction Region 2; the interaction region in which BaBar is located
IRSAMC	Institut de Recherche sur les Systèmes Atomiques et Moléculaires Complexes
ISR	Initial State Radiation
IT	Information Technology
JADE	JADE detector at DESY stands for Japan, Deutschland and England
JFY	Japanese Fiscal Year
JLab	Thomas Jefferson National Accelerator Facility
JRA	Joint Research Activity
JSON	JavaScript Object Notation
KEK	High Energy Accelerator Research Organisation
KVM	Kernel-based Virtual Machine is a virtualization infrastructure for the Linux kernel
L3	High Energy Physics Experiment at the LEP collider
LAL	Laboratoire de l'Accélérateur Linéaire
LCP	Laboratoire de Physique Corpusculaire
LEP	Large Electron Positron Collider
LHC	Large Hadron Collider
LHCb	LHC-beauty; an LHC experiment
LPCNO	Laboratoire de Physique et Chimie des Nano-Objets
LRI	Laboratoire de Recherche en Informatique
LSF	Load Sharing Facility
LSST	Large Synoptic Survey Telescope
LTDA	Long Term Data Access
MatLab	A numerical computing environment commercialized by MathWorks
MC	Monte Carlo
MDST	Mini Data Summary Tape
MGT	Management of the consortium
mODS	Micro Object Data Store
MPS	Multiparticle Spectrometer facility at the BNL AGS
MSSM	Minimal Supersymmetric Standard Model
MSST	Massive Storage Systems and Technology
NASA	National Aeronautics and Space Administration
NASA-ADS	The Astrophysics Data System (usually referred to as ADS), developed by the National Aeronautics and Space Administration (NASA), is an online database of over eight million astronomy and physics papers from both peer reviewed and non-peer reviewed sources
NDB	H1 database software package
NEUROBA	An advanced neural network implementation

YES	
NFS	Network File System
NLO	Next to Leading Order
NNLO	Next-to-next-to-leading order
NSF	National Science Foundation
NVO	National Virtual Observatory
OAIS	Open Archival Information System
OCR	Optical Character Recognition
ODS	Object Data Store
ODE	Opportunity for Data Exchange
OPAL	Omni-Purpose Apparatus for LEP
OPENAIRE/ OPENAIRE PLUS	Open Access Infrastructure for Research in Europe
OPR	Online Prompt Reconstruction
OS	Operating System
OSG	Open Science Grid
PARSE. Insight	Permanent Access to the Records of Science in Europe
PARTNER	Particle Training Network for European Radiotherapy
PAW	Physics Analysis Workstation is an interactive graphical data analysis program
PB	Peta Byte
PBS	Portable Batch System
PC	Project Coordinator
PEP-II	Accelerator at SLAC National Accelerator Laboratory
PETRA	Positron-Elektron-Tandem-Ring-Anlage
PM	Project Month
PMB	Project Management Board
PubDb	BaBar Publications Database System
PWA	Partial Wave Analysis Techniques
QCD	Quantum Chromodynamics
QED	Quantum electrodynamics
QR	Quarterly Report
R&D	Research and Development
RAID	Redundant Array of Independent Disks (originally Redundant Array of Inexpensive Disks) is a storage technology that provides increased reliability and functions through redundancy
RAL	Rutherford Appleton Laboratory
RAW	Unprocessed data direct from the detector
RECAST	A framework to fully exploit the power of existing physics analyses to guide the community it its search for new physics
RHEL	Red Hat Enterprise Linux
ROOT	An object-oriented program and library developed by CERN
ROSCOE	Robust Scientific Communities for EGI
SAM	A data handling system at DØ
SARA	Stichting Academisch Rekencentrum Amsterdam

SCOAP3	Sponsoring Consortium for Open Access Publishing in Particle Physics
SDSS	Sloan Digital Sky Survey
SL	Scientific Linux is a Linux Operating System based on RHEL
SLAC	SLAC National Accelerator Laboratory (formerly Stanford Linear Accelerator Center)
SLD	Scientific Linux DESY
SuperB	High-luminosity electron-positron collider that will be dedicated to elucidating new physics through precision studies of rare or suppressed decays
SUSY	Supersymmetry
TAO	The Ace Orb is a freely available, open-source, and standards-compliant real-time C++ implementation of CORBA based upon the Adaptive Communication Environment (ACE)
TB	Tera Byte
TDS	Transient Data Store
Tevatron	Tevatron particle collider, at the Fermilab in Batavia, Illinois, USA, so named because the energy of each beam reach 1 TeV
Twiki	An Open Source Enterprise Wiki
TF	Technical Forum
TMB	Technical Management Board
ULICE	Union of Light Ion Centres in Europe
UVIC	University of Victoria, Canada
VM	Virtual Machine
VO	Virtual Organisation
VRC	Virtual Research Communities
WLCG	Worldwide LHC Computing Grid
WP	Work Package
WWW	World Wide Web
Xen	The Xen hypervisor is a powerful open source industry standard for virtualization
XFER	Transfer
XROOTD	The XROOTD project aims at giving high performance, scalable fault tolerant access to data repositories of many kinds
ZEUS	Particle detector on the HERA particle accelerator at DESY

8. Annex: Letters of Support

Table 4 – Letters of Support for the ILDAP Proposal

Person	Position
Michael Ernst	Director of RHIC and ATLAS Computing Facility, Brookhaven National Laboratory (BNL), US.
Victoria White	Associate Director for Computing Science and Technology, Chief Information Officer, Fermi National Accelerator Laboratory (FNAL), US.
Guido Tonelli	Spokesperson, CMS experiment, INFN and the University of Pisa, Italy
Ruth Pordes	Executive Director, Open Science Grid (OSG), US, Associate Head of the Fermilab Computing Sector (FNAL), US.

BROOKHAVEN
NATIONAL LABORATORY

Physics Department
P.O. Box 5000
Upton, NY 11973-5000
Phone 631 344-4755
mernst@bnl.gov

managed by Brookhaven Science Associates
for the U.S. Department of Energy

www.bnl.gov

November 8, 2011

Dr. Jamie Shiers
Information Technology Department
CERN

Dear Dr. Shiers,

I as the Director of the RHIC and the ATLAS Computing Facility (RACF) at Brookhaven National Laboratory (BNL) and US ATLAS Facility Manager am writing to you in support of the "International Long-term Data and Analysis Preservation (ILDAP)" project proposal that was submitted by CERN, CNRS and DESY as participating institutions in response to the INFRA-2012-3.2 program call.
The ATLAS Computing Facility at BNL is the largest out of ten Tier-1 centers worldwide supporting analysis of data taken with the ATLAS detector at the Large Hadron Collider (LHC) at CERN. The RACF is a shared facility that also serves as the main data archive and analysis center for the nuclear physics program performed at the Relativistic Heavy Ion Collider (RHIC) operated at BNL. PHENIX and STAR, the main experiments at RHIC with more than 500 collaborators each from twelve countries add currently more than 3 PB each year to the data archive. Both programs together have currently at BNL an active data volume of 15 PB that is expected to grow to at least 50 PB by the end of this decade.

We see several specific scenarios where the preservation of experimental particle and nuclear physics data would be of benefit to the respective communities: An extension of the existing physics program may be necessary to ensure the long term completion of ongoing analysis ; it may be favorable to re-do previous measurements to achieve an increased precision: reduced systematic errors may be possible via new and improved theoretical calculations (MC models) or newly developed analysis techniques; preserving old data sets may allow the possibility to make new measurements at energies and processes where no other data exists. Finally, if new phenomena are found in new data at the LHC or some other future collider, it may be useful or even mandatory to go back, if possible, and verify such results using older data.

Given the tremendous value of data obtained at detectors over many years, if not decades, long term international coordination in the area of data preservation in High Energy and Nuclear Physics is essential. While a scientific supervision of the preserved data sets is considered as mandatory, coordination on the international scene will ensure a coherent and extensive usage of the potential of the preserved data sets. It will also enforce the persistence of various data sets against possible local resource problems. Investments in local data preservation programs are therefore enhanced by an international organization.

We look forward to continuing our international collaboration to benefit the research community, and in enabling the capabilities to be further extended and used.

Please feel free to contact us if there is any additional information you may need.

Sincerely,

Michael Ernst
Director, RHIC and ATLAS Computing Facility,
U.S. ATLAS Facility Manager
Brookhaven National Laboratory
Upton, New York, USA

Fermilab

Fermi National Accelerator Laboratory
Victoria A. White
Associate Lab Director for
Computing Science and Technology
and Chief Information Officer
MS370 * P.O. Box 500 * Batavia, IL 60510
Office 630/840-3936 * Fax: 630 840 3785
Email white@fnal.gov Cell: 630 774 9552

November 11, 2011

Dr. Jamie Shiers
CERN
CH-1211 GENEVE 23
Switzerland

Subject: Fermilab support for the International long-term Data and Analysis Preservation (ILDAP) Project

Dear Jamie:

On behalf of the Fermilab National Accelerator Laboratory we are writing in strong support of the ILDAP project proposal.

Fermilab is the largest US national laboratory devoted to particle physics research. The scientific mission and program of the laboratory encompass research based on the highest energies achievable at particle accelerators and high intensity beams used to study rare and interesting phenomenon. A large and vibrant program of particle astrophysics that studies the structure and nature of the universe, with an emphasis on dark matter and dark energy are also major components of the mission of the laboratory.

The areas of science Fermilab studies are very data intensive over long timescales. Data is collected for many years and sizable samples are studied and analyzed well past the end of data taking. Fermilab has many experiments that have collected large datasets that are currently being analyzed. These include the just-concluded Tevatron Run 2 experiments CDF and D0, with dataset volumes of approximately 10 PBytes each. Fermilab is the largest Tier 1 computing facility for the CERN CMS experiment and has responsibilities for storage, processing and analysis of CMS data. Currently Fermilab stores approximately 13 PBytes of data associated with CMS. Other experiments associated with the Fermilab scientific program store substantial datasets as well.

The ILDAP project to study the long-term availability and sharing of large datasets is a particularly important issue at this time. The work that is to be accomplished in this program will move the field forward and will allow for further possibilities of collaboration and research. This is particularly interesting and relevant to Fermilab because of the recent ending of the Tevatron Run 2 data taking. The question of longer-term data analysis of those datasets is an important issue for the laboratory and for the field of particle physics. Fermilab is working together the Run II experiments during 2012 on initial activities useful to the long-term availability of the Run II datasets. We look forward to collaborating with ILDAP on those that are mutually relevant across the international EU and US physics groups of the collaborations.

The ILDAP project will be beneficial to the laboratory, to the world-wide particle physics community and to the larger science community that is interested in learning how to make productive use of the data sets.

Sincerely,

Victoria A. White
Associate Director for Computing Science and Technology
Chief Information Officer
Fermi National Accelerator Laboratory

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH
COMPACT MUON SOLENOID COLLABORATION

URL : http://cms.cern.ch

Prof. Guido Emilio Tonelli
CMS Spokesperson
CERN – PH Department
CH - 1211 GENEVA 23

Tel. *+41 22 767 1477*
Fax *+41 22 767 8940*
E-mail *Guido.Tonelli@pi.infn.it*

Dr. Jamie SHIERS
Information Technology
Department
CERN

CH-1211 Genève 23

Geneva, November 16, 2011

Votre référence / Your reference :
Notre référence / Our reference : CMS-20111116/GET/ka

Subject: *Data preservation*

Dear Dr. Shiers,

I strongly support your proposal entitled "International Long-term Data and Analysis Preservation (ILDAP)" that was submitted by CERN, CNRS and DESY in response to the INFRA-2012-3.2 program call.

The LHC data is extremely valuable and will be of interest to the community for decades to come. Already, at the end of year 2 of the first run, the CMS archive alone is measured in tens of petabytes of storage and this will continue to increase rapidly in the years to come. There are more than a thousand active analyzers of the CMS data and a tremendous amount of accumulated expertise.

Effective data preservation has historically been a challenge to the field and is an activity that has frequently been left until late in an experimental program. Coordinating the effort across experiments and institutions, as outlined in your proposal, will increase the value of the effort and starting this activity early will increase the chances of success. We believe that augmenting the effort in a coherent program for data preservation will benefit the CMS community specifically and the research community in general.

CMS looks forward to continuous international collaboration in this important area and believes that your proposal is an interesting opportunity to move forward.

Yours Sincerely,

Prof. Guido Emilio Tonelli
CMS Spokesperson
CERN, Geneva (Switzerland)
University and INFN, Pisa (Italy)

*Adresse postale pour le courrier posté en France : CERN : Site de Prévessin, F-01631 CERN Cedex

Letter of Commitment from Open Science Grid to collaborate with the International Long-term Data and Analysis Preservation (ILDAP) Project
November 18th 2011

Dear Jamie,

The Open Science Grid (OSG) Consortium provides a multi-disciplinary collaboration in support of domain science communities distributed computing and data needs across DOE and NSF in the United States. The US LHC communities are among our key stakeholders. Other physics and astro-physics communities are members of the Consortium and users of the operations, software and support services provided. The OSG works together with the US LHC collaborations on the use and support of those resources and services of the World Wide LHC Computing Grid located in the US.

Many of the current OSG stakeholder communities plan to reuse, provide open access, curation and preservation of their scientific data as well as sustain, verify and validate the software environments used for the analysis of the data over the long term. We plan to propose and work collaboratively with ILDAP on prototyping these capabilities and understanding the processes and agreements that need to be put in place. We expect such a collaboration to be through proposal of an OSG Satellite project – an independent research and development project that contributes to the overall mission of the OSG and its stakeholders.

Clearly through collaboration with ILDAP we will work on capabilities and prototypes needed by the LHC experiments. We would plan to also work with our other communities, especially those in Astrophysics, on technology and procedural areas that can be leveraged and/or in common. We would deliver software developed to support these capabilities through the OSG Virtual Data Toolkit.

To reflect the equal importance of the data itself and the software used to access and process it, we are currently referring to this plan as the Data and Software Preservation for Open Science (DASPOS).

We look forward to continuing our international collaboration with you on these very important capabilities needed by our scientific stakeholders.

Sincerely

Ruth Pordes
OSG Executive Director