

# ILDAP

## *International Long-term Data and Analysis Preservation*

V1.06  
16 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      |

**Comment [s1]:** *I think we're forgetting this part. It is very important to write a 2-3 pages description on why we are doing the international co-operation and how.*

*Some bullet points you could use:*

- HEP is global by construction*
- Preservation is a global problem and "owners" of the data set (those who paid and those who took them) are mixed group of DoE, NSF and European funding agencies*
- We have buddies who're going to work with us see letters of support*
- We've hands-on projects in DPHEP already with DESY but also SLAC (get letter of support for their preservation project)*
- We use platforms as INSPIRE for preserving documentation/metadata which are jointly operated by CERN-DESY and DOE-funded SLAC and Fermilab*
- Say that there are projects at NSF which are going to work with us and ask for money to do so*

## Table of Contents

|        |   |    |
|--------|---|----|
| 1.     | Scientific and technical quality.....                                   | 3  |
| 1.1.   | ILDAP Concepts and Objectives .....                                     | 3  |
| 1.2.   | ILDAP Coordination Overview .....                                       | 9  |
| 1.3.   | ILDAP Workplan.....   | 11 |
| 1.3.1. | Risks and Contingency Plans.....  | 24 |
| 2.     | Implementation.....   | 26 |
| 2.1.   | Management structure and procedures.....                                | 26 |
| 2.1.1. | Project coordinator .....   | 26 |
| 2.1.2. | Coordinating Partner.....   | 26 |
| 2.1.3. | Administrative Assistant.....   | 26 |
| 2.1.4. | Project Management Board .....  | 26 |
| 2.1.5. | Technical Management Board .....  | 27 |
| 2.2.   | Individual participants.....  | 28 |
| 2.2.1. | CERN.....   | 28 |
| 2.2.2. | CNRS .....  | 29 |
| 2.2.3. | DESY .....  | 30 |
| 2.3.   | Consortium as a whole.....  | 31 |
| 2.4.   | Resources to be committed .....   | 33 |
| 3.     | Impact.....   | 34 |
| 3.1.   | Expected impacts listed in the work programme .....                     | 34 |
| 3.2.   | Spreading excellence, exploiting results, disseminating knowledge ..... | 37 |
| 3.2.1. | Project Internal.....   | 37 |
| 3.2.2. | External Dissemination and Outreach.....                                | 37 |
| 3.2.3. | Educational Outreach.....   | 38 |
| 3.2.4. | Policy Makers and Future Strategies / Work .....                        | 38 |
| 4.     | Ethical Issues .....  | 40 |
| 5.     | Annex: References.....  | 41 |
| 6.     | Annex: Glossary .....   | 42 |
| 7.     | Annex: Letters of Support .....   | 48 |

## 1. Scientific and technical quality

### 1.1. ILDAP Concepts and Objectives

#### 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 in the context of ongoing 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 in a large-scale survey of the HEP community performed as part of the PARSE.Insight Support Action, also funded under this program. [<http://arxiv.org/abs/0906.0485> and [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-3\\_CaseStudiesReport.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf)] In order to perform this work a coordination body would be established that would not only organize workshops devoted to this topic but 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 the both during and after the lifetime of the corresponding scientific collaboration, considering opportunities offered by Open Data concepts weighted against the community practices and self-organising collaborations. Moreover, networking of experts both within the HEP domain but also 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.

#### Long-term completion and extension of scientific programs

This entails the natural continuation of the physics program 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

**Comment [s2]:** *I see all the concepts in the following text. Nice and clear.*

*I DO NOT UNDERSTAND WHICH ARE THE OBJECTIVES !!!*

*Maybe some bullet points with three clear SMART objectives would help the evaluator finding them (and giving the proposal more points for clarity in that part)*

**Comment [jds3]:** *To be addressed later today...*

## ILDAP

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.

### 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 search of new physics. Preserved data sets may further enhance the physics potential of experimental programs, 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, 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.

**Comment [s4]:** *There is too much jargon here... shall we define things? Do we assume physicists are going to be the evaluators? I would not do so!*

**Comment [jds5]:** *To be addressed later today...*

### 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 the data preservation has not been planned in most of the experiments, examples of the usefulness of a 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

**Comment [s6]:** *If the evaluator is not a physicist, s/he will not understand what all this is about. I would strongly suggest to try and have a pedestrian explanation of all these concepts. You can leave maybe in a footnote all the details. But the risk we have here already is this is already a very "vertical" single-discipline proposal, so do not make it worse with all this physics.*

**Comment [jds7]:** *OK - delete or put in appendix and add simple example.*

## ILDAP

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. [JADE]

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. It is the case 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 the models involving the so-called dark photons. These bosons would result from a special theory extending the 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 the neutral pions to photons. These branching ratios are best measured in the so-called beam-dump experiments, performed for the essential 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 the 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]

In conclusion, the persistence of data analysis in HEP can and has led to new scientific opportunities. The exposed examples may well be only the top of the iceberg. The paradigm itself and the addressed issues (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 of data samples explosion (see astrophysics or life sciences) HEP will need an data analysis approach that is closer to other sciences and may play a leading role in building synergies for the future Experimental Data Observatories.

Furthermore, 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 as it is and has been in other fields – 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 use of 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 the currently ongoing and planned studies, the LHC data – being very rich – will have a large physics potential even after the active data-taking period.

Given the long life-time of the LHC experiments and the large volume of the collected data, the data preservation has to be addressed already during the active data-taking. The LHC experiments can take advantage of the experience of the previous experiments' data preservation activities and apply timely the measures ensuring data preservation. 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 worldwide. The LHC experiments started with a fully distributed environment where the vast majority of the resources are located away from CERN. The LHC Computing Grid was approved by CERN Council in 2001 and it evolved into the Worldwide LCG (WLCG) with service support for all 4 LHC experiments.

The LHC experiments will need to address the risk of loss of data due to obsolescence of enabling technologies and due to physical damage. The risk due to physical damage is largely covered by the distributed storage in professional computing centres. The threat of obsolescence of hardware and software environment will require proactive measures ensuring that the data files will remain readable and usable in the long-term future.

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 redone after the LHC long shut-down foreseen 2013-2014. Lessons learnt from such exercises will be incorporated in the long term preservation of the data and associated software.

The details of the data preservation plan after the data-taking will not be defined at this early stage, but the LHC experiments will follow with attention the procedures taken by the experiments which have recently ended the data taking or are in the final analysis phase. This experience will be useful for proactive planning of the long-term future.

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 publishing and storing them at external, persistent repositories. In addition to the written article, additional public data sets such as numerical values of the tables can be provided broadening the concept of the scientific publication. This is already being experimented and INSPIRE [INSPIRE] is planned to be the long-term platform for such additional information. Common efforts between experiments and theorists can be made more efficient if data are presented in a way that they can be combined and compared either with other experiments or with theoretical predictions. This will ensure a vaster public re-use of the scientific data by the emerging community of citizen scientists, enticed by the media attention to the LHC program.

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

**Comment [s8]:** *This is not entirely correct, as instead the LHC is starting to prepare them. I would say so. And say that they're doing so as instruments as the "NSF data management plan" are creating an opportunity for the global LHC collaborations to do so. So ILDAP is a great opportunity to bring together challenges from previous experiments, more regionally funded, and these requests on a global level, and lay the groundwork for a collaboration with the US (and others) in this area of the e-Infrastructure*

**Comment [jds9R8]:** *Will do!*

## ILDAP

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 organization 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 programs targeting high school students. The text-based ".ig" file format is human-readable and largely self-explanatory. It uses the JSON standard format meaning it is easy to read programmatically with C++, Python, etc. without the need for any CMS-specific software. The use of common simplified formats for open access will be explored.

The LHC experiments are currently defining their data preservation and access policies and plans. A common policy statement for all the experiments will be based on the levels of the data preservation model in the DPHEP context. Each experiment will provide a plan how the policy will be implemented in the experiment-specific context. The challenges are very similar across experiments, and some nuances could exist at the access level, as a function of time. At the moment, the current practices are being examined to evaluate what is needed to extend them to a functional long-term data preservation model. This proposal will play a unique role in harmonizing the policies of the LHC experiments, the technologies and requirements being developed by hands-on preservation of other, existing, data sets, and the wider background of emerging standards and trends in data preservation, such as those analysed in the APARSEN Network of Excellence, also funded by this directorate.

### **Education, training and outreach**

Preserving data opens new opportunities in training, education, and outreach. It permits data analysis by undergraduate or graduate students, without restriction to institutes that collaborated to the experiments, opening 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 schools 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, co-ordinated by CERN, also funded under this same scheme.

## 1.2. ILDAP Coordination Overview

HEP is an international discipline that spans the globe. Scientists are organized into *collaborations* that are associated (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 and 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 can be expected to subsequently be of significant benefit also to other communities.

**Comment [s10]:** *I do not understand if this is something that ILDAP will do or something behind the economic and sociologic necessity of collaboration to emerge as self-organising structures.*

Exactly in the same way in which building and operating detectors, as well as analysing data and publishing articles is a global endeavor, the preservation of data is a global issue: no individual team working within a collaboration, or a single laboratory, can command the resources needed for such an effort. A participative process is key. At the same time, the funding and organizational mechanisms in the field are ill-conceived for open-ended efforts dedicated to data-sets which are not emerging from an operating facility. Therefore a new, global, co-ordination structure has to be designed, with both the scientific mandate of supervision of the preserved data set, as well as the maximization of access opportunity and, therefore, scientific potential. Such a global infrastructure will inherently insure the persistence of various data sets against possible local resources problems. At the same time, investments in local data preservation programs are therefore enhanced by an international organization.

An analysis of the scientific potential of the preserved data can be made for experiments approaching the end of the scientific program (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 been observed as well 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. This simple counting is enhanced, as explained above, by the fact that some of the subjects may become critical or even crucial if new findings point to possible re-analysis and confirmations using unique data sets.

**Comment [s11]:** *I think we should put here the graph of Cristi where one sees three plots, the luminosity going up fast. Then the money disappearing also fast. Then the people evaporating... and show then also the papers continuing on this which is the best data.*

The costs of dedicated preservation programs (2-3 FTEs over a few years) are in fact very small when compared to the investments and represent of the order of 1/1000 of the costs of construction and running (not including the full scientific person power). These programs can have the maximal output if an international coordination is installed, such that a critical mass is achieved for the activity and an enhancement of the collaborations and data usage is steered on the international scene and with a long term perspective. In addition, common dedicated projects, as the ones proposed in this project, are made possible via international, multi-experimental and multi-laboratory cooperation.

**Comment [s12]:** *All these FTEs here are very confusing. The evaluators will not understand if those are the one you ask in this proposal or what. I propose to remove the FTEs and use this as JUST THE RECCOMANDATIONS about what has to be done.*

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

## ILDAP

catastrophic loss of data once major collaborations come to an end. The recent expertise gained during the last 18 months indicate that an extension of the computing effort within experiments with a person power of the order of 2-3 FTEs leads to a significant improvement in the ability to move into a long-term data preservation phase. Such initiatives exist already or are being defined in the participating laboratories and are followed attentively by the Study Group.

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

**Priority 3: Common R&D projects.** Common requirements on data preservation are likely to evolve into inter-experimental R&D projects (three concrete examples are given above, each involving 1-2 dedicated FTE, across several laboratories). The projects will optimise the development effort and have the potential to improve the degree of standardisation in HEP computing in the longer term. Concrete requests will be formulated 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 the lessons learnt by the PARSE.Insight and ODE support actions financed by this same scheme, running from 2008 to 2012. Through a series of multi-disciplinary international workshops, this group identified two main use cases: educational and for 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 summarizing and eventually extending the requirements capturing 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 involving 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, also 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

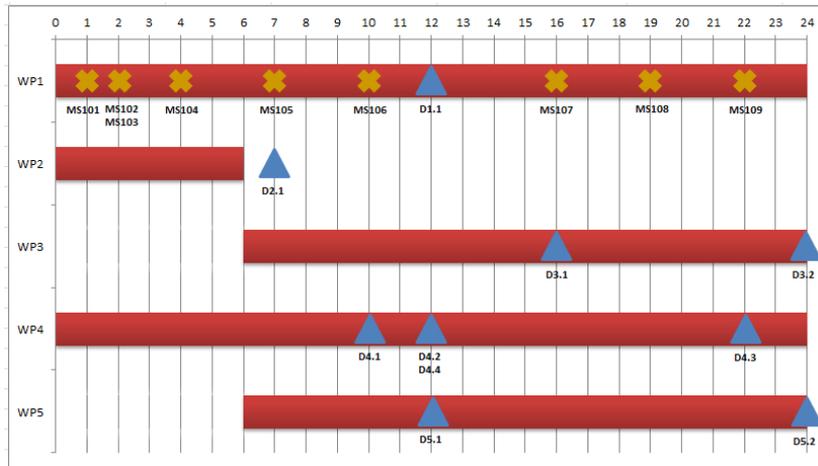
The overall strategy is to build on existing work, in particular that performed by the DPHEP consortium, and take it to the next stage, establishing first a demonstrator and then a prototype of a system that supports long-term data preservation and enables re-analysis of the data. As such there are preparatory 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. This will continue to strength collaboration within the HEP community as well as further discussions with other communities facing similar problems and undertake the all important tasks related to dissemination of the results, outreach.

**Comment [s13]:** I think we have a small confusion on the word outreach which we use here for both what the proposal teaches the world and when HEP goes to schools. I would use "dissemination" for the first and "outreach" for the latter. It might need sanitizing throughout.

## ILDAP

### 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 (standardization) and 5 (prototypes).



#### Key



Milestone



Deliverable

#### List of Milestones

MS101 – Project Mailing lists  
MS102 – Project Templates  
MS103 – Website

#### List of Deliverables

D1.1 – Project Annual Report  
**D1.2 – Project Final Report** (missing in Gantt chart above)  
D2.1 – Requirements Summary  
D3.1 – Interim Report on Standardisation Activities  
D3.2 – Final Report on Standardisation Activities  
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

**Comment [s14]:** In the Gantt this is at M12, should be M24

**Table 1.3 a: Work package list**

| Work package No | Work package title | Type of activity | Lead participant No | Lead participant short name | Person-months | Start month | End month |
|-----------------|--------------------|------------------|---------------------|-----------------------------|---------------|-------------|-----------|
| WP1             | Project Management | MGT              | 1                   | CERN                        | 24            | 1           | 24        |
| WP2             | Requirements       | COORD            | 3                   | DESY                        | 10            | 1           | 6         |
| WP3             | Standardisation    | COORD            | 2                   | CNRS                        | 38            | 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 Annual Report                                  | 1      | R      | PU                  | M12           |
| D1.2     | 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      | O      | PU                  | M12 M9?       |
| D5.2     | Prototype  | 5      | O      | PU                  | M24 M21?      |

**Comment [s15]:** Don't you want to have this at M9 (in pre-release maybe as another deliverable) to show at the first workshop?

**Comment [s16]:** As above, if you had something earlier, for the workshop, as well would make lots of sense.

**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 |
| MS103            | Website               | WP1                      | M2            | Project website setup and in use                                       |
| MS109            | Quarterly Report 6    | WP1                      | M22           | Quarterly Report Submitted   |

**Comment [s17]:** *Make it a deliverable as well... nice and cheap!*

**Comment [s18]:**

*If you want more milestones, use the pre-release of the prototype/demonstrators ready for presentation at the first and second workshop (2 nice new milestones)*

*And add the transfer of information from WP2 to WP3 and WP5*

**Table 1.3 d: Work package description**

|                                       |                    |                                      |     |
|---------------------------------------|--------------------|--------------------------------------|-----|
| <b>Work package number</b>            | WP1                | <b>Start date or starting event:</b> | PM1 |
| <b>Work package title</b>             | Project Management |                                      |     |
| <b>Activity Type</b>                  | MGT                |                                      |     |
| <b>Participant number</b>             | 1                  |                                      |     |
| <b>Participant short name</b>         | CERN               |                                      |     |
| <b>Person-months per participant:</b> | 24                 |                                      |     |

**Comment [s19]:** I think you might want to have 0.5-1 PM from each of the other partners as well for mere participation to phone calls and meetings

**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**

Given the size of the consortium and project the management will be kept as light-weight as possible (see section 2.1). In particular, it will be responsible for:

**Comment [s20]:** Maybe this then does not justify even 24PM/150PM to do it?

- 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: Annual Report describing the progress made by the project during the first year. (PM12).
- D1.2: Final Report describing the progress made by the project during the second year. (PM24).

**Table 1.3 d: Work package description**

|                                       |              |                                      |    |  |  |  |  |
|---------------------------------------|--------------|--------------------------------------|----|--|--|--|--|
| <b>Work package number</b>            | WP2          | <b>Start date or starting event:</b> | M1 |  |  |  |  |
| <b>Work package title</b>             | Requirements |                                      |    |  |  |  |  |
| <b>Activity Type</b>                  | COORD        |                                      |    |  |  |  |  |
| <b>Participant number</b>             | 3            | 1                                    |    |  |  |  |  |
| <b>Participant short name</b>         | DESY         | CERN                                 |    |  |  |  |  |
| <b>Person-months per participant:</b> | 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 be made available 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.

ILDAP

**Table 1.3 d: Work package description**

|                                       |                 |                                      |      |  |    |
|---------------------------------------|-----------------|--------------------------------------|------|--|----|
| <b>Work package number</b>            | WP3             | <b>Start date or starting event:</b> |      |  | M7 |
| <b>Work package title</b>             | Standardisation |                                      |      |  |    |
| <b>Activity Type</b>                  | COORD           |                                      |      |  |    |
| <b>Participant number</b>             | 2               | 2                                    | 1    |  |    |
| <b>Participant short name</b>         | CNRS            | DESY                                 | CERN |  |    |
| <b>Person-months per participant:</b> | 18              | 12                                   | 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**

Select or if necessary define the standards to preserve the analysis capability for long periods of time by migrating to the latest technologies and software versions for as long as possible, substantially extending the lifetime of the software and thus, of the data.

This work will be carried out in close collaboration with the host laboratories and research institutes in Europe as well as their counterparts in the US (e.g. FNAL for the Tevatron, SLAC for BaBar and BNL for RHIC).

**Comment [s21]:** *This seems a bit thin a description for 38 PM! I think you've to write much more!*

**Comment [jds22R21]:** *CNRS needs to write this IMHO as WP leader.*

**Deliverables**

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

**Table 1.3 d: Work package description**

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

**Objectives**

- Harmonise and synchronise preservation projects across all stakeholders and collaborate with relevant initiatives from other fields
- Talk to experiments, primarily in HEP but also across multiple scientific domains, and stress the advantages of their involvement
- Organize workshops to ~~foment~~ **enlarge the community involved in the preservation ties.** the involvement of the different experiments
- ~~Publish the results in conferences and journals~~
- Extend and explore connections with a large spectrum of experiments in HEP
- Define the principles, the operational model and the agreements needed for an international organization centered on the DPHEP proposal.

**Comment [s23]:** *This does not make sense. Do we speak out of HEP for experiments to teach us something or for them to learn something? I think you need two sentences and clarify the different things we want from the two different target audiences.*

**Comment [s24]:** *This is a byproduct of all the activity. I think you ought to remove it.*

**Formatted:** Strikethrough

**Formatted:** Strikethrough

**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 will cover the main goals identified by DPHEP, which are listed below.

- 4-1) Position itself as the natural forum for the entire discipline to foster discussion, achieve consensus, and transfer knowledge in two main areas
  - 4-a) Technological challenges in data preservation in HEP
  - 2-b) Diverse governance at the collaboration and community level for served data
- 2-2) Co-ordinate common R&D projects aiming to establish common discipline-wise preservation tools
- 3-3) Harmonize preservation projects across all stakeholders and liaise with relevant initiatives from other fields
- 4-4) Design the long-term organization of sustainable and economic preservation in HEP
- 5-5) Outreach within the community and advocacy towards the main stakeholders for the case of preservation in HEP

**Comment [s25]:** *Isn't this the same of point 2 above, talking with experiments? This is confusing.*

**Comment [s26]:** *I am not really sure this is a job of networking. I think networking allows you to learn all this but the work gets done somewhere else. If you do it in less than 7 months you do it in requirements... IF not, maybe in standards?*

**Formatted:** Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

**Formatted:** Outline numbered + Level: 2 + Numbering Style: a, b, c, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Tab after: 0.5" + Indent at: 0.5", Tab stops: Not at 1"

**Formatted:** Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

**Deliverables**

Other: International Workshops, involving EU, US and others

ILDAP

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

**Table 1.3 d: Work package description**

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

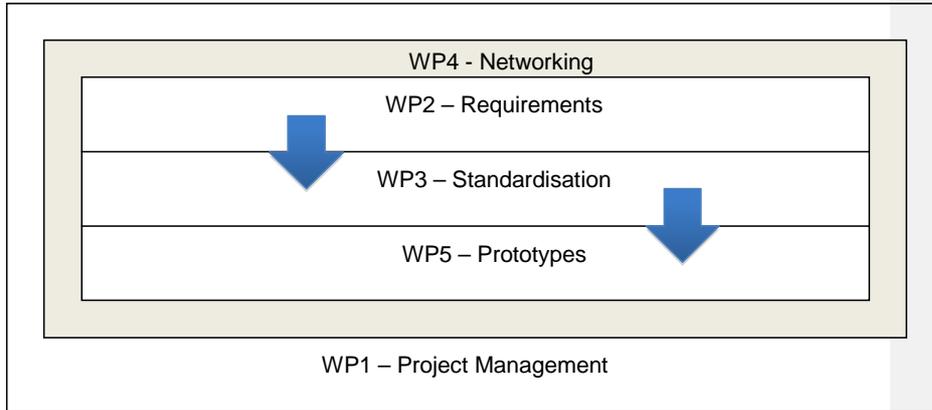
A Demonstrator of solutions to the technological problems will be provided by PM12, followed by a Prototype system in PM24.

ILDAP

**Table 1.3 e: Summary of staff effort**

| <b>Participant no./short name</b> | <b>WP1</b> | <b>WP2</b> | <b>WP3</b> | <b>WP4</b> | <b>WP5</b> | <b>Total person months</b> |
|-----------------------------------|------------|------------|------------|------------|------------|----------------------------|
| 1 CERN                            | <b>24</b>  | 4          | 8          | <b>18</b>  | 18         | 72                         |
| 2 CNRS                            |            |            | <b>18</b>  |            | 18         | 36                         |
| 3 DESY                            |            | <b>6</b>   | 12         | 12         | <b>18</b>  | 48                         |
| Total                             | 24         | 10         | 38         | 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**

| <b>Work Package</b> | <b>Risk</b>                         | <b>Impact</b>   | <b>Probability</b> | <b>Mitigation</b>  |
|---------------------|-------------------------------------|---|--------------------|--|
| 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 has 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 organization 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 Administrative Assistant in daily administrative and financial management.

These bodies would hold regular phone or video conferences in addition to technical meetings organized 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.

ILDAP

## 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 plethora 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 barrier 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 to Open Access the entire HEP literature. 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.

### 2.2.2. CNRS

#### Brief description of legal entity

The Centre National de la Recherche Scientifique (National Centre for Scientific Research) is a government-funded research organisation, under the administrative authority of France's Ministry of Research. CNRS's 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 Programmes. Over the past years, the CNRS has acquired an outstanding experience in coordinating FP Projects.

#### Main tasks in the project and relevant experience

The Laboratory involved in the LDAP project is CC-IN2P3, the National Computing Centre of IN2P3 Institute. 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 High Energy Physics. More than 2000 users working on 40 international experiments use its services. As a Tier1 for 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.

Since 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 has graduated in the field of experimental High energy physics participated as physicist in all the fields of a running experiment, from detector operation to software and analysis of the data. She has participated to experiments at CERN as well as in the USA, at Fermilab. 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 finalized. 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 to as per the work package breakdown tables in section 1.

ILDAP

### 2.2.3. DESY

#### **Brief description of the legal entity**

The “Stiftung Deutsches Elektronen-Synchrotron 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 organization supported by public funds.

#### **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 High Energy Physics 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 3<sup>rd</sup> 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 long-term data preservation for future analyses. Both CERN and DESY have considerable experience in the data management and storage domain as both 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<sup>1</sup> 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, 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 re-analysis 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 reaching 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 barrier 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

---

<sup>1</sup> Other solutions include StoRM, BeStMan and xrootd.

## ILDAP

(WP4). Over the last few years DESY has participated in the DPHEP organisation, 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 (24PM), 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 (132PM). 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 organization 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 programs in the HEP scientific community. Natural continuation of the programs 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 groundbreaking 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 programs 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 program. 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 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 programs 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 thanks to a standardisation of the format of high-level data, 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, archival infrastructures. Solutions for the automatic migration to new media generation and technology, as well as for the automatic data integrity checks will be designed, that will also bring a 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  $e^+e^-$  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 anal-

## ILDAP

ysis 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 the data preservation would allow this kind of analysis to be done without extra efforts.

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 standardization in order to define the working groups. We plan to involve all partners, including those without Data Preservation program, which will benefit from such project in the future.
- Month 6 - A technical forum to discuss the possible paths for Standardisation and Prototypes. The TF should coincide with the beginning of the Standardisation and Intellectual Property Framework work packages.
- 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 18 – Technical Forum to check point on the various activities and prototypes
- 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 Symposium (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).

ILDAP

### 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 visualization, 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 interested public
2. High school science students
3. College students studying courses in particle physics or computing
4. Graduate students in particle physics.

Each group provides different challenges and each may interact with these data in different ways.

Thus this work will raise the overall awareness and appreciation of the scientific work so that it becomes an integral part of future educational models addressing all of 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 potential 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. Similar, data in the Life Science domain can be expected to grow significantly as more patient information is made available digitally, such as scans of various sorts. Thus, the work supported by this project can be seen as valuable input into preparing longer term and more diverse

## ILDAP

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 |
|---|-----|------|
| <b>Informed Consent</b>   |     |      |
| • 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?  |     |      |
| <b>Research on Human embryo/foetus</b>  |     |      |
| • Does the proposal involve Human Embryos?  |     |      |
| • Does the proposal involve Human Foetal Tissue / Cells?  |     |      |
| • Does the proposal involve Human Embryonic Stem Cells?   |     |      |
| <b>Privacy</b>  |     |      |
| • 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?   |     |      |
| <b>Research on Animals</b>  |     |      |
| • 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?   |     |      |
| <b>Research Involving Developing Countries</b>  |     |      |
| • Use of local resources (genetic, animal, plant etc)   |     |      |
| • Impact on local community   |     |      |
| <b>Dual Use</b>   |     |      |
| • Research having direct military application   |     |      |
| • Research having the potential for terrorist abuse   |     |      |
| <b>ICT Implants</b>   |     |      |
| • Does the proposal involve clinical trials of ICT implants?  |     |      |
| <b>I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL</b>   | YES |      |

## 5. Annex: References

Table 3 – References

|                      |   |
|----------------------|---|
| <b>DPHEP</b>         | ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics – <a href="http://www.dphep.org">http://www.dphep.org</a> . |
| <b>JADE</b>          |   |
| <b>BLEUMLEIN</b>     |   |
| <b>INSPIRE</b>       | Production service at <a href="http://Inspirehep.net">http://Inspirehep.net</a>   |
| <b>PARSE.Insight</b> | Permanent Access to the Records of Science in Europe<br><a href="http://www.parse-insight.eu/">http://www.parse-insight.eu/</a>                 |
| <b>ODE</b>           | Opportunity for Data Exchange <a href="http://ode-project.eu">http://ode-project.eu</a>   |
| <b>APARSEN</b>       | Alliance for Permanent Access to the Records of Science Network<br><a href="http://aparsen.eu">http://aparsen.eu</a>                            |

## 6. 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       | <u><a href="#">Belle Analysis Framework</a></u>   |
| 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     | <u><a href="#">CERN Advanced Storage Manager</a></u>  |
| CCIN2P3    | <u><a href="#">Centre de Calcul de l'IN2P3</a></u>  |
| 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 Organization 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  |
| ConditionD<br>B       | 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<br>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                   | <a href="#">Digital Object Identifier System</a>  |
| DPHEP                 | Study Group for Data Preservation in High Energy Physics  |
| DPM                   | <a href="#">Disk Pool Manager</a>   |
| 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-<br>InSPIRE       | EGI Integrated Sustainable Pan-European Infrastructure for Re-searchers in Europe                                   |
| EnviroGRID<br>S       | 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<br>XFEL      | ESFRI project: X-ray Free Electron Laser research infrastructure  |

ILDAP

|          |   |
|----------|---|
| 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  |
| GDML     | <u>Geometry Description Markup Language</u>   |
| GEANT    | Detector Description and Simulation Tool  |
| GeV      | Gigaelectron Volt   |
| GKS      | Graphical Kernel System is the first ISO standard for low-level computer graphics   |
| GLAST    | <u>Fermi Gamma-ray Space Telescope, formerly the Gamma-ray Large Area Space Telescope</u>   |
| GPD      | Generalized Parton Densities  |
| GridKa   | <u>Grid Computing Centre Karlsruhe</u>  |
| 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 loosely 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     | <u>High Performance Storage Systems</u>   |
| HSM      | Hierarchical Storage Management   |
| HTML     | <u>HyperText Markup Language</u>  |
| 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     | <u>Institute of High Energy Physics</u>   |
| 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             | <u>Information Technology</u>  |
| 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 Organization  |
| 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<br>YES | An advanced neural network implementation  |
| NFS            | Network File System  |

ILDAP

|                   |   |
|-------------------|---|
| 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  |
| OPR               | Online Prompt Reconstruction  |
| OS                | Operating System  |
| OSG               | Open Science Grid   |
| PARSE.<br>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 in 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   |
| 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   |
| 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  |

ILDAP

## 7. Annex: Letters of Support

**Table 4 – Letters of Support for the ILDAP Proposal**

| <b>Person</b>  | <b>Position</b>   |
|----------------|---|
| 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. |
| 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  
memat@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 (ACF) 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 ACF 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.

ILDAP

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

ATLAS  
Brookhaven National Laboratory

11/11/2011

11/11/2011

11/11/2011

11/11/2011

Dear [Name],

I am writing to you regarding the ATLAS computing facility. We are currently looking for a Director, RHIC and ATLAS Computing Facility, U.S. ATLAS Facility Manager. This is a full-time position and will be based at Brookhaven National Laboratory, Upton, New York, USA. The position involves the management of the ATLAS computing facility and the coordination of the ATLAS computing resources. The successful candidate will have a Ph.D. in a related field and a minimum of 10 years of experience in the field of high energy physics computing. The successful candidate will also have a strong background in the management of large scale computing systems and a strong leadership ability. If you are interested in this position, please send your resume and cover letter to [Email Address].

We are an equal opportunity employer. Minorities and women are encouraged to apply. If you are a minority individual who believes you have been discriminated against in any aspect of your employment, you may file a complaint with the EEOC. For more information, please contact the EEOC at 1-800-368-1024.

Thank you for your interest in this position. We look forward to receiving your application.

Sincerely,  
Michael Ernst



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](mailto: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 encompasses 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