Table of Contents

INFRA - 2012 - 3.2 Project Proposal Preparation

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

INFRA-2012-3.2:

The objective is to establish an EU/USA coordination platform aiming at full interoperability of scientific data infrastructures, and to demonstrate this coordination through several joint EU-USA prototypes that would ensure persistent availability and effective sharing of data across scientific domains, organisations and national boundaries. The platform should provide for: the collection of requirements and approaches for standardisation (development, promotion, adoption and maintenance); common ICT infrastructure approaches (technical, semantic, reference architecture, financing models, etc) in order to lower access barriers; harmonisation of intellectual property frameworks for scientific information; and mechanisms for international networking of experts and multidisciplinary communities. The joint prototypes should leverage and build upon similar initiatives in Europe and USA. The proposal should clearly describe synergies and collaboration with corresponding existing or potential NSF-funded initiatives.

Status of Proposal Preparation

- SubmittedFiles

Part	Responsible	Comments	Status
A1	Jamie Shiers	Draft validated in EPSS	Draft
A2.CERN	Jamie Shiers	Checked by CERN EU office, uploaded to EPSS	Closed
A2.CNRS	Ghita Rahal	Checked by CNRS EU office, Uploaded to EPSS	Open
A2.DESY	Volker Guelzow	Checked by DESY EU office, Uploaded(by DESY EU office) to EPSS	Closed
A3.CERN			
A3.CNRS			
A3.DESY			

PART A

Draft A1 Summary Form

- **Proposal Acronym:**
  - ILDAP - International Long-term Data and Analysis Preservation

- Other Proposed Acronyms
  - PreSciDatIAnl = PREservation of SCIentific DATa for International ANalysis Long-term (pronounced "presi-day-tional")
  - SciDP = Scientific Data Preservation
  - IDEAS = International Data Exchange Across Sciences
  - PreScieNT = Preservation of Scientific data - Novel Techniques (yuk!)
  - PreSciEUS
  - POISED = Preservation Of International Scientific Experiment Data
  - DESciPLE = Data Exchange & Scientific Preservation for Long-term Experiments
  - PreSciLA = PREservation of SCIentific data and/for Long term Analysis
  - PreSciLLA = PREservation of SCIentific data and/for Long Lived Analysis
  - ICoPAD/ICPAD = International Coordination for the Preservation of Analysis Data
  - TACoPAD/TACPAD = Trans-Atlantic Coordination for the Preservation of Analysis Data

- PADICA = Preserving Analysis Data (through) International Coordination Activities
- CAPAD = Coordinating Activities for the Preservation of Analysis Data
- ADaPtICS = Analysis Data Preservation through International Coordination Strategies
- ICADaPt = International Coordination of Analysis Data Preservation Technology/Techniques
- PreSciENT = Preservation of Scientific Experimental
- PreDICE = Preservation of Data by International Coordination Efforts
- CAPED = Coordinating Activities for the Preservation of Experimental Data
- PreScienCE = Preservation of Scientific data through Coordinated Effort
- PreScICE = Preservation of Scientific data through Internationally Coordinated Effort (Pronounced precise)
- HEDAP High energy data and analysis preservation
- SDAP Scientific Data and Analysis Preservation
- hepPAD high energy physics preservationof analysis and data
- DPHEP (just use what people know already)
- INELAPSED - INternational Effort for the Long-term Analysis (Availability) and Preservation of Scientific Expertise and Data
- UNELAPSED - UNited Effort for the Long-term Analysis (Availability) and Preservation of Scientific Expertise and Data
- DAPLE = Data and Analysis Preservation for Long-term Exploitation
- DAPLES = Data and Analysis Preservation for Long-term Experimental Science

- **Proposal Title:**
  - Preservation of Scientific Data for International Analysis Long-Term (or Long-Term Analysis)
  - Common Interoperable International Infrastructures for the Preservation of Scientific Data for Long-Term Use
  - (Or simply (an expansion of) the acronym?)

- **Duration in months:**
  - The suggestion is for a 24 month project which will overlap with the start of FP8 / Horizon 2020 (2014 on)

- **Call (part) Identifier:** FP7-INFRASTRUCTURES-2012-1

- **Topic code(s) most relevant to you proposal:** *Refer to the topic codes/objectives listed in the work programme call fiche. All activities and topics of FP7 have been assigned unique codes, which are used in the processing of data on proposals and subsequent contracts. The codes are organised hierarchically. The choice if the first topic code will be limited in the drop-down menu to one of the topic open in this call. Select the code corresponding to the topic most relevant to your proposal. The choice for the second code is also limited to topics open in the call in question. Enter a second code if your proposal also addresses another of these. Select none if this is not the case. Select a third code if your proposal is also relevant to another theme. This time, the available codes will simply correspond to the broad themes. Select none if this is not the case.*

- **Free Keywords:**
  - Scientific Data Preservation Archive EU US NSF

- **Abstract:**
  - The preservation of scientific data for long-term analysis has been identified as a key requirement in the field of High Energy Physics (HEP) and also in many other disciplines such as Astronomy and Astrophysics, Life and Earth Sciences. In collaboration with related projects in the US and in close collaboration with the National Science Foundation (NSF), the proposed project would take the work of the Data Preservation in HEP Study group, which defines the physics motivation for long-term data preservation as well as many of the

associated issues, and extend this to cover not only the existing use cases but also consider the 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. In particular, now is the ideal time to define standards for data and meta-data formats and address access and authorization issues for on-going experiments issues that have historically been addressed only in the final years of a scientific collaboration. In order to perform this work a coordination body would be setup that would not only organize workshops devoted to this topic but also address key issues related to long-term data archives, such as the financing models for maintaining these archives, the handling of intellectual property rights both during and after the lifetime of the corresponding scientific collaboration, as well as the required networking of experts both within the HEP domain but also with other disciplines and projects, such as those mentioned above, with whom close collaboration already exists. The results of this work would be made available via Open Access mechanisms and would be 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 and Medical Imaging Symposium and other similar events.

- **Similar Proposals or signed contracts:**
  - None (AFAIK)

Draft A2 Participant Forms

- CERN A2 Form
- DESY A2 Form (outdated)
- To follow

Useful Links

- A1 Summary Form
- A2 Participants Form
- Talk at EGI-T Lyon September 2011
- EU Participant Portal: http://ec.europa.eu/research/participants/portal/page/capacities?callIdentifier=FP7-INFRASTRUCTURES-201
- ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics: https://www.dphep.org/

PART B

Section Documents

- Cover Page & TOC
- Section 1 - Scientific/Technical Quality
- Section 2 - Implementation
- Section 3 - Impact
- Section 4 - Ethics

Section Structure and Content

- **1. Scientific and/or technical quality, relevant to the topics addressed by the call.** *20 pages for the whole section. (Not including charts/diagrams/tables)*
- **1.1 Concept and Objectives** Explain the concept of your project. What are the main ideas that led you to propose this work? *No specific limit.*

- **1.2 Contribution to the coordination of high quality research** Indicate how the area addressed by your project will benefit from the co-ordination (including networking) that you propose. *No specific limit.*
- **1.3 Quality and effectiveness of the co-ordination of high quality research** A detailed work plan should be presented, broken down into work packages (which should follow logical phases of the implementation of the project) and include consortium management and assessment of progress and results. *1 page for section 1.3 (i) ("Overall strategy")*
- **2 Implementation**
- **2.1 Management structure and procedures** Describe the organisational structure and decision-making mechanisms of the project. Show how they are matched to the complexity and scale of the project. *5 pages*
- **2.2 Individual participants** For each participant in the proposed project, provide a brief description of the legal entity, the main tasks they have been attributed, and the previous experience relevant to those tasks. Provide also a short profile of the staff members who will be undertaking the work. *1 page per participant*

This is the description of CERN used in the ROSCOE proposal:

Brief description of legal entity: CERN, the European Organization for Nuclear Research, is the largest particle physics labor-atory in the world and is an International Organisation with its headquarters in Switzerland. Currently CERN is commissioning the Large Hadron Collider (LHC), a new particle accel-erator on the Swiss-French border near Geneva, expected to be operational at the end of 2009. LHC is the world s most powerful accelerator and will provide research facilities for several thousand high-energy physics 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. These experiments will run up to 1 million computing tasks per day and will generate around 15 petabytes of data per year. This data will be shared with all the participating institutes. The computing capacity required to analyse the data far exceeds the capacity needs of any comparable physics experiments today and relies on the combined re-sources of some 200 computer centres world-wide. CERN and the particle physics communi-ty have chosen grid technology to address the huge data storage and analysis challenge of LHC. Main tasks in project and relevant experience CERN will be involved in work packages NA2, NA3 as well as SA1, SA2 and SA3. IT department of CERN currently has 228 staff, predominantly engineers, who operate one of Europe s largest research computer centres supporting about 17,000 users. The department has developed leading expertise in large-scale data centres and long-standing collaborations with industrial and academic partners in the fields of high performance computing and ad-vanced networking. The CERN IT department has been at the forefront of computing for many years and now leads the world s largest grid project, EGEE (Enabling Grids for E-SciencE). CERN has also prominently contributed to a number of EGEE-related grid projects aiming at extending the EGEE production grid infrastructure to new geographical areas, to serve new applications domains and to support the grid community: BalticGrid-II, D4Science, D4Science-II, EGI_DS, enviroGRIDS, ETICS 2, GridTalk, Health-e-Child and SEE-GRID-SCI. Under FP6 and FP7, the department has been involved in some 20 European Commission-funded pro-jects. CERN is a founding partner of the recently formed European Grid Initiative that will provide a sustainable grid infrastructure for Europe s research communities. Profiles of individuals undertaking the work

Dr Jamie Shiers currently leads the Grid Support group in CERN s IT department. He has been involved in grids since the early days of the European Data Grid. Since 2005 he led the WLCG Service Challenge activity and later organized the Common Computing Readiness Challenge 2008 and the Scale Test for the Experiment Programme 2009 two key demon-strations of production readiness of WLCG.

Jakub T. Moscicki is a software engineer and researcher at CERN. He obtained the MSc in Computing at the AGH University of Science and Technology in Krakow. His research and engineering interests focus on the distributed and parallel applications deployed on large-scale computing infrastructures such as the Grids. He is the technical leader within the Ganga project (and the creator of DIANE) now used by several hundred users from High-Energy physics and other sciences. Maarten Litmaath has more than 20 years of in-depth

experience with POSIX-compliant computing and UNIX derivatives. He joined the CERN IT department in 2002 and since then has been working on building up the Worldwide LHC Computing Grid (WLCG) as well as the Enabling Grids for E-sciencE (EGEE) infrastructure.

- **2.3 Consortium as a whole** Describe how the participants collectively constitute a consortium capable of achieving the project objectives, and how they are suited and are committed to the tasks assigned to them. Show the complementarity between participants. Explain how the composition of the consortium is well-balanced in relation to the objectives of the project. *No specific limit*
- **2.4 Resources to be committed** Describe how the totality of the necessary resources will be mobilised, including any resources that will complement the EU contribution. Show how the resources will be integrated in a coherent way, and show how the overall financial plan for the project is adequate. Please also mention any other major costs that haven't been previously mentioned (e.g. equipment). *2 pages*

- **3 Impact** *10 pages for whole section*
- **3.1 Expected impacts listed in the work programme** Describe how your project will contribute towards the expected impacts listed in the work programme in relation to the topic or topics in question. Mention the steps that will be needed to bring about these impacts. Explain why this contribution requires a European (rather than national or local) approach. Indicate how account is taken of other national or international research activities. Mention any assumptions and external factors that may determine whether the impacts will be achieved.
- **3.2 Spreading excellence, exploiting results, disseminating knowledge** Describe the measures you propose for the dissemination and/or exploitation of project results, and how these will increases the impact of the project. In designing these measures, you should take into account a variety of communication means and target groups as appropriate (e.g. policy-makers, interest groups, media and the public at large).

- **4 Ethics Issues** Describe any ethics issues that may arise in the project. In particular, you should explain the benefit and burden of their experiments and the effects it may have on the research participants. All countries where research will be undertaken should be identified. You should be aware of the legal framework that is applicable and the possible specific conditions that are relevant in each country (EU and non-EU countries alike). It is strongly advised that when drafting the research proposal, the local ethics committee or/and relevant competent authorities should be contacted for information and, when applicable, guidance. *No limit*

For the full instructions for Part B see the "Useful Links" section below.

Adaptation of the ROSCOE Proposal

- Part B - Sections 1-8 Adapted from the ROSCOE Proposal
- Section 4 - Ethical Issues
- Section 5 - Gender Action Plan
- Section 6 - References
- Section 7 - Glossary
- Section 8 - Letters of Support

Useful Links

- Part B Instructions
- Eligibility and evaluation criteria

Agendas

- Meeting of Oct 26 2011

Mailing list

- Archives

-- ThomasMannifield - 20-Oct-2011

This topic: LCG > ProjectPreparation
Topic revision: r50 - 2011-11-23 - JamieShiers