Open access and data preservation in H2020

Silvia Amerio
(Università di Padova & INFN)

Outline

- *Introduction on open access (OA) and data preservation (DP)*
- *Activities outside and inside HEP*
- *H2020 calls*
- *Proposals*

Introduction

Data preservation (DP)

Preserve the capability to access and analyse **all data** collected by an experiment

Bit preservation

Software preservation

Documentation

Open Access (OA)

Data accessible to general public (which data can be made public? How can general public access the data? How can they analyse it?....)

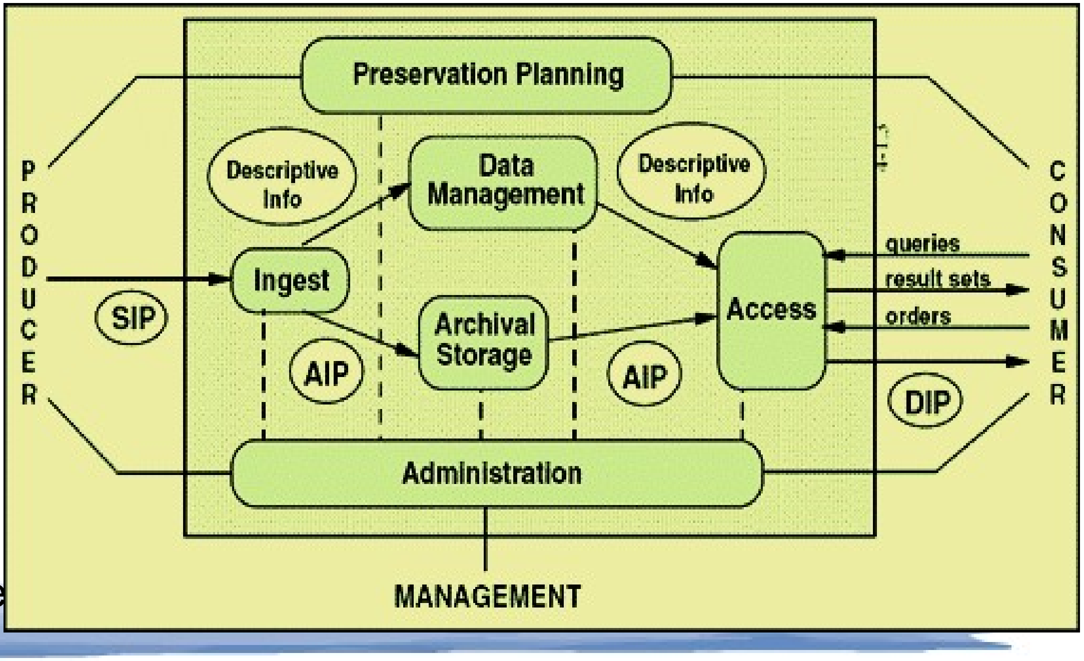
Interest in the ***long term preservation of scientific data and their availability to general public*** *is growing; governments and funding agencies are recommending it.*

Some examples from outside HEP

Astronomy is well ahead in data preservation thanks to standard formats and open access to data.

FITS (Nasa)

Flexible Image Transport System (FITS) is an open standard defining a digital file format useful for storage, transmission and processing of scientific and other images. FITS is the most commonly used digital file format in astronomy



OAIS (Esa)

Reference model for an open archival information system

DP in HEP

HEP is still behind, each experiment has its own data format, software, data access policies, etc..., but we are trying to narrow the gap:

Past experiments have developed/are developing DP plans (Babar, DESY, TEVATRON)
All **LHC experiments** are devoting more efforts to data preservation (dedicated task forces, regular meetings, ...)

DPHEP (Data Preservation in High Energy Physics)

Study group started in 2009, blueprint released in May 2012 (available at dphep.org)
Priorities:

- *Secure data in all experiments*
- *Consolidate the on-going international collaboration*
- *Promote common multi-experiment projects and/within interdisciplinary cooperation*

DPHEP vision:
By 2020 all archived data easily findable, fully usable by designated communities with clear (open) access policies.
Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards.

DP projects in HEP

BaBar Long Term Data Access system installed for analysis until at least 2018. Isolated from SLAC, it uses *virtualisation techniques to preserve and existing stable and validated platform.*

New development of BaBar LTDA: BaBar-To- Go. A virtual machine, in raw format, that can run with multiple software applications on multiple platforms (KVM, VMware, VirtualBox, …) , with one or more BaBar analysis releases fully installed and ready to use (Talk by T.Cartaro here at CHEP2013).

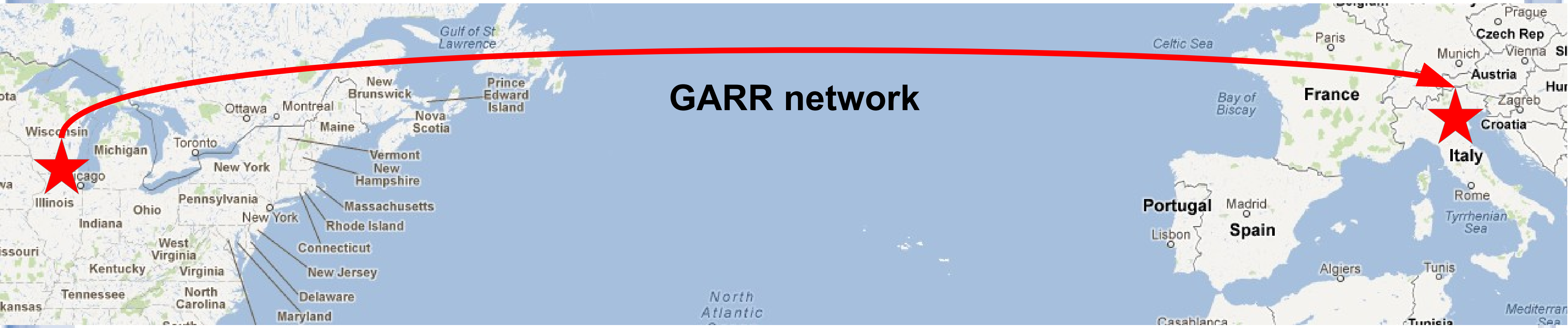
Aleph has deployed a SL4 virtual machine working in Virtual Box that is used by INFN Bari cloud service (poster at CHEP2013 here)

Desy: Generic validation framework for data analysis in HEP experiments, to keep the experimental software up-to-date with changes of the OS (poster at CHEP2013 here)

CDF/D0: run 2 data preservation project to ensure at least until 2020 data access and analysis: migration of data to new tape technology, legacy software release, virtualisation techniques to run experiment software, Inspire to archive all documentation.

INFN-CNAF ltdp project

Goal: preserve a complete copy of CDF data and MC samples at CNAF + services (access, data analysis capabilities)



The copy will be split in two years

- end 2013 - early 2014 → All data and MC user level ntuples (2.1 PB)
- mid 2014 → All raw data (1.9 PB) + DBs

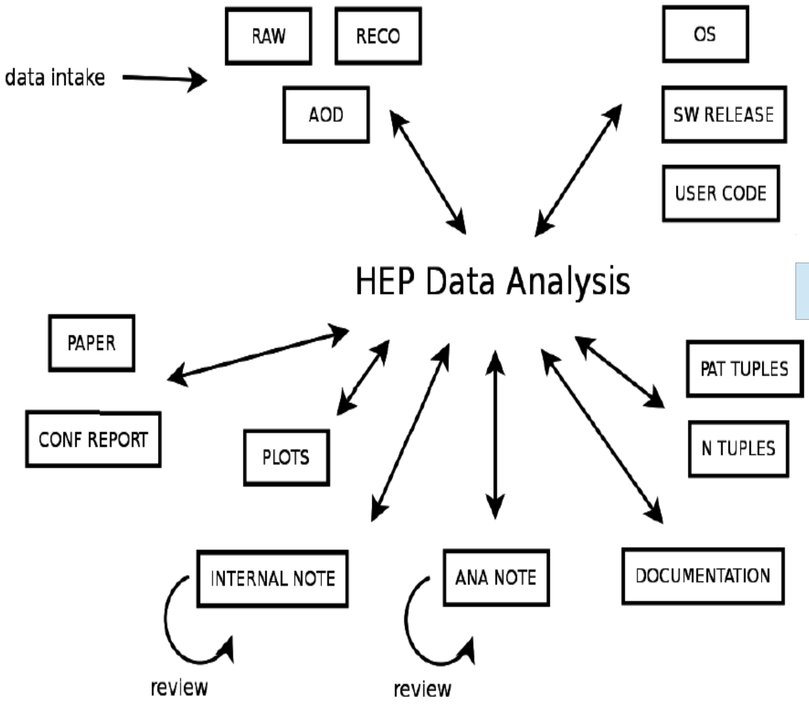
During the second year (2014), development of the long term future analysis framework, based on WNoDes.

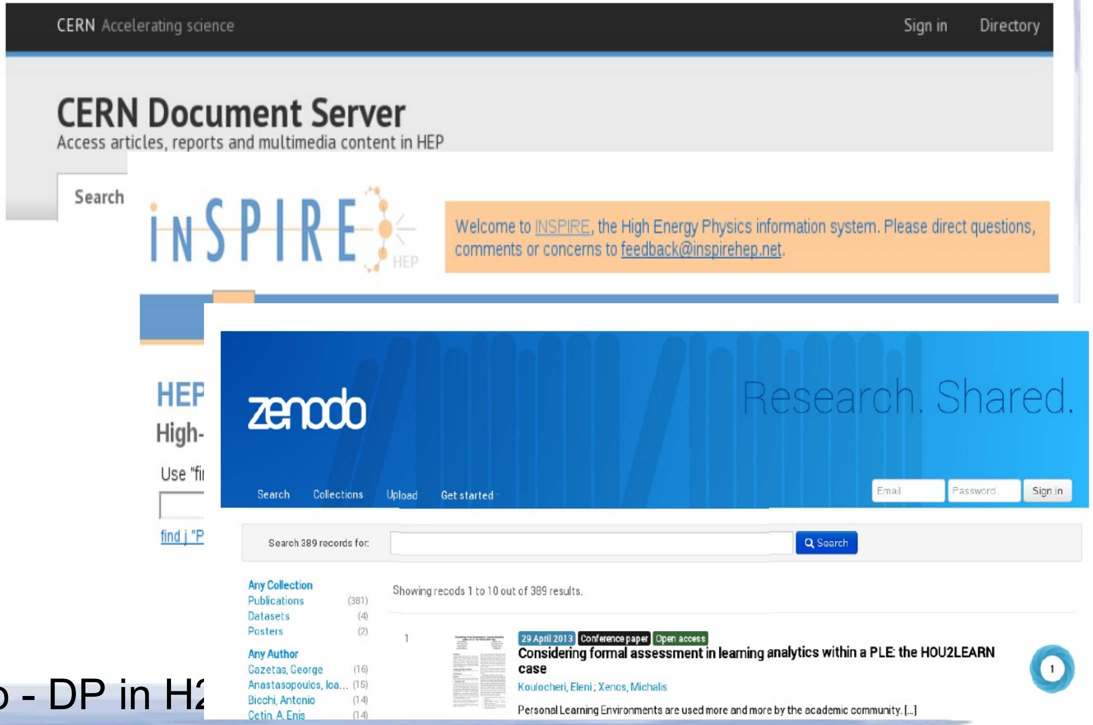
- Preserve data access
- Preserve CDF reconstruction and analysis software
- Give users resources to run CDF analysis (authentication, disk space, CPU)
- Documentation

DP @ LHC

CMS/LHCb

- open access policy approved by the collaboration (CMS will release their data in 2014, LHCb in 2018)
- definition of legacy datasets and legacy software releases
- development of long term future validation frameworks
- The Invenio team at CERN, with input from CMS and LHCb, will prototype extensions of CDS make recording and documenting of the workflow and intermediate data easy





CERN Document Server

Access articles, reports and multimedia content in HEP

INSPIRE HEP

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

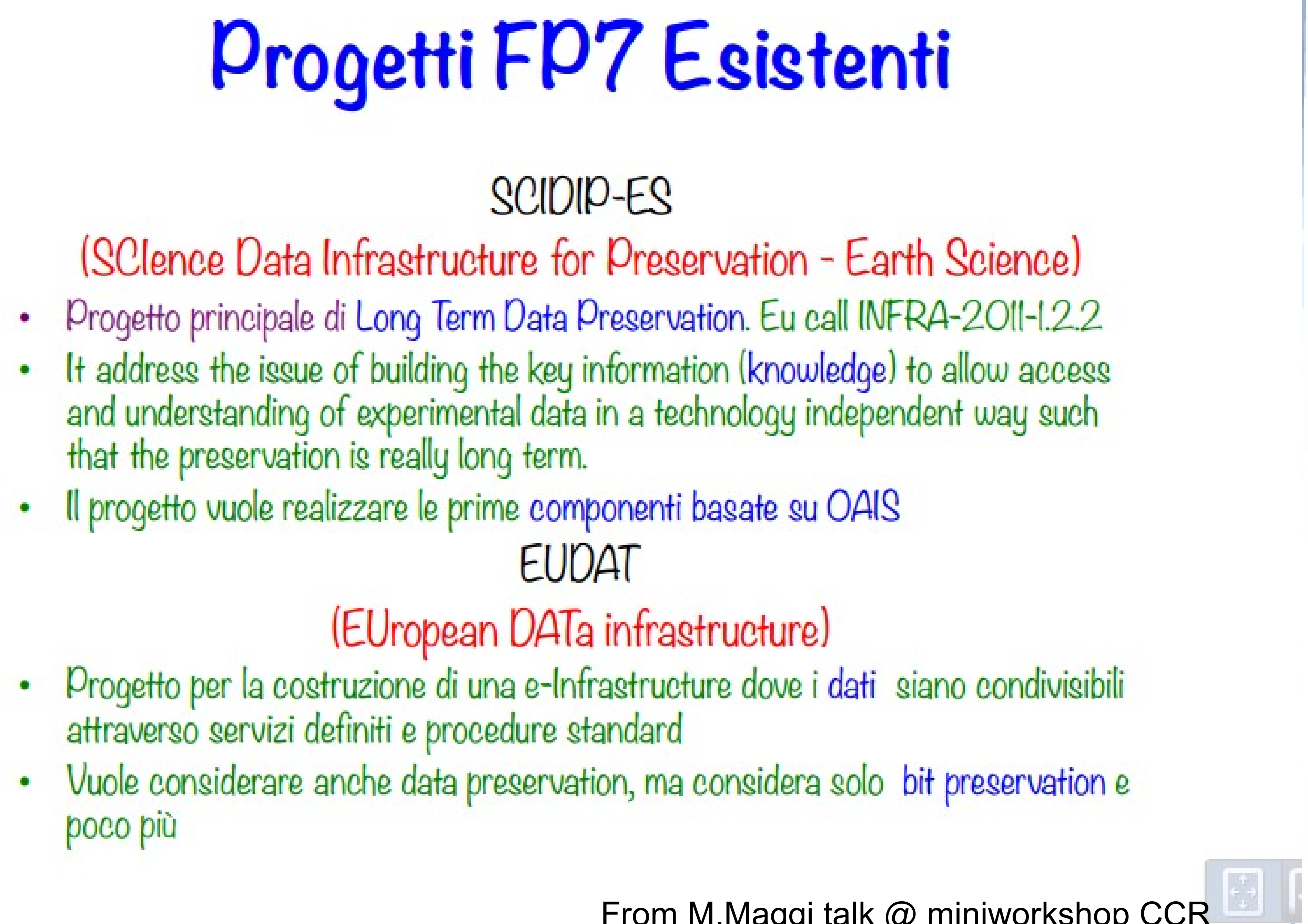
Search 389 records for:

Showing recods 1 to 10 out of 389 results.

Considering formal assessment in learning analytics within a PLE: the HOU2LEARN case

Koulocheri, Eleni ; Xenos, Michalis

Personal Learning Environments are used more and more by the academic community. [...]



Progetti FP7 Esistenti

SCIDIP-ES

(SCIence Data Infrastructure for Preservation - Earth Science)

- Progetto principale di Long Term Data Preservation. Eu call INFRA-2011-1.2.2
- It address the issue of building the key information (knowledge) to allow access and understanding of experimental data in a technology independent way such that the preservation is really long term.
- Il progetto vuole realizzare le prime componenti basate su OAIS

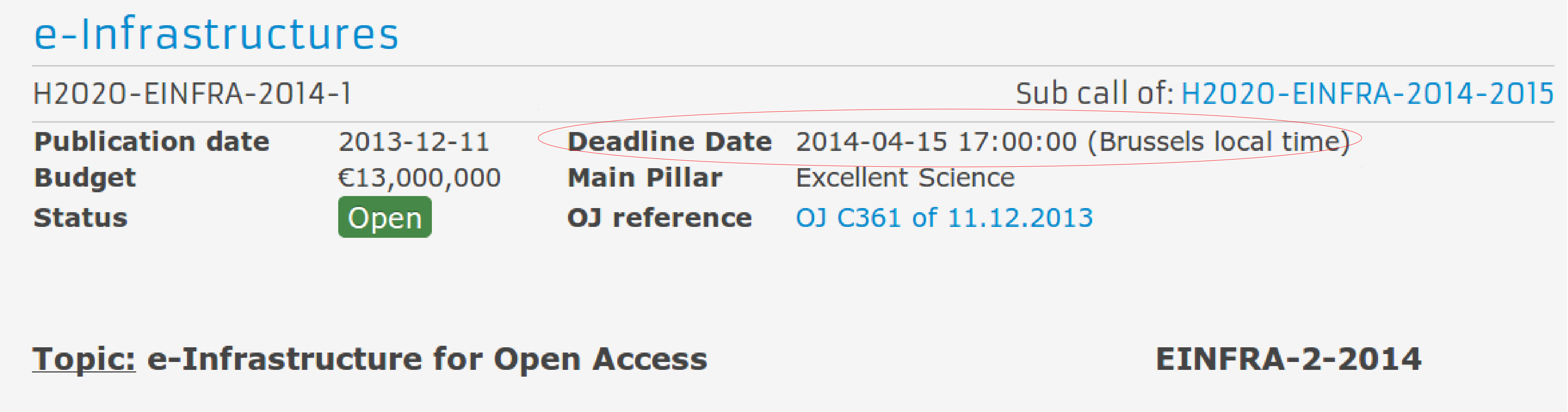
EUDAT

(EUropean DATa infrastructure)

- Progetto per la costruzione di una e-Infrastructure dove i dati siano condivisibili attraverso servizi definiti e procedure standard
- Vuole considerare anche data preservation, ma considera solo bit preservation e poco più

From M.Maggi talk @ miniworkshop CCR

H2020 calls on OA and DP



e-Infrastructures

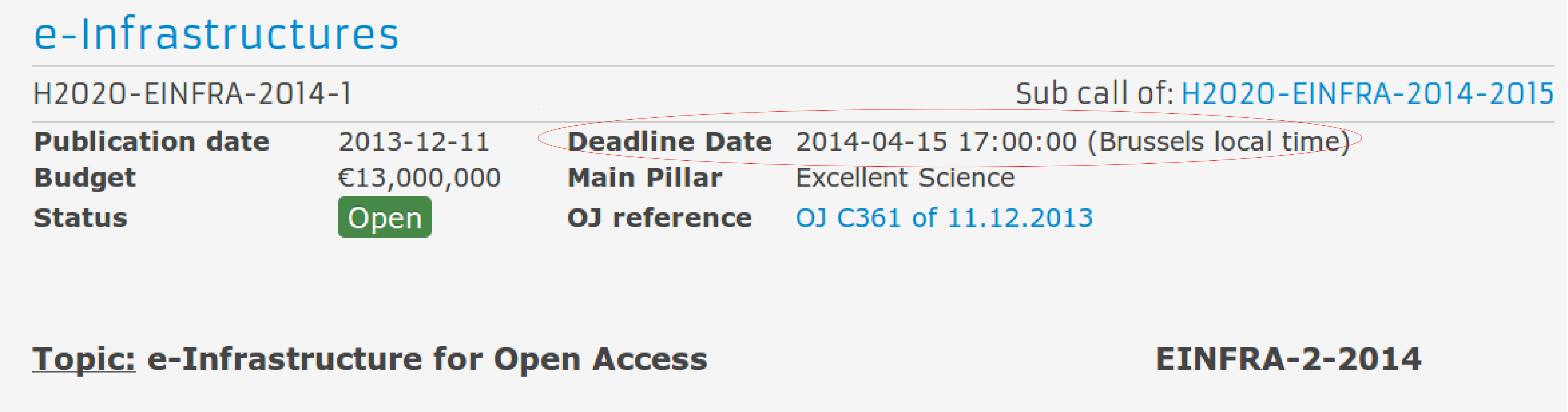
H2020-EINFRA-2014-1 Sub call of: H2020-EINFRA-2014-2015

Publication date	2013-12-11	**Deadline Date**	2014-04-15 17:00:00 (Brussels local time)
Budget	€13,000,000	**Main Pillar**	Excellent Science
Status	Open	**OJ reference**	OJ C361 of 11.12.2013

Topic: e-Infrastructure for Open Access **EINFRA-2-2014**

Specific challenge: Europe needs a robust **e-infrastructure supporting Open Access policies**, also for Horizon 2020. This infrastructure, **based on already existing e-infrastructures** (institutional and thematic repositories, aggregators, etc.), should support reliable and permanent access to digital scientific records. **A key element will be capacity building to link literature and data in order to enable a more transparent evaluation of research and reproducibility of results.** Such an action will include an analysis of alternative means of public support to Gold Open Access in order to identify the optimal approach. The Open Access mandate and the Open Data Pilot of Horizon 2020 impose new requirements for the infrastructures to fully support participants to comply with their obligations and objectives. Therefore, a key objective will be to provide service driven infrastructures to enable wide participation in the Open Data Pilot.

H2020 calls on OA and DP



e-Infrastructures

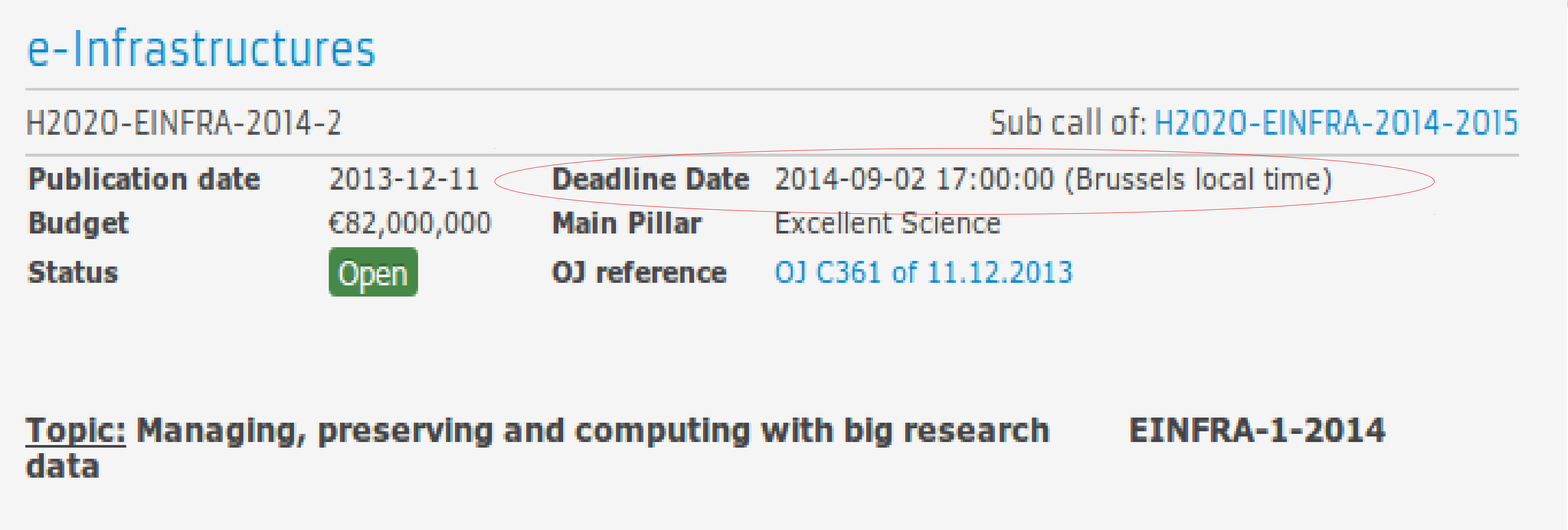
H2020-EINFRA-2014-1 — Sub call of: H2020-EINFRA-2014-2015

Publication date	2013-12-11	**Deadline Date**	2014-04-15 17:00:00 (Brussels local time)
Budget	€13,000,000	**Main Pillar**	Excellent Science
Status	Open	**OJ reference**	OJ C361 of 11.12.2013

Topic: e-Infrastructure for Open Access **EINFRA-2-2014**

Specific challenge: Europe needs a robust **e-infrastructure supporting Open Access policies**, also for Horizon 2020. This infrastructure, **based on already existing e-infrastructures** (institutional and thematic repositories, aggregators, etc.), should support reliable and permanent access to digital scientific records. **A key element will be capacity building to link literature and data in order to enable a more transparent evaluation of research and reproducibility of results.** Such an action will include an analysis of alternative means of public support to Gold Open Access in order to identify the optimal approach. The Open Access mandate and the Open Data Pilot of Horizon 2020 impose new requirements for the infrastructures to fully support particip e to provide

Possibilità di inserimento di INFN nel progetto Cern-IT-CMS-LHCb



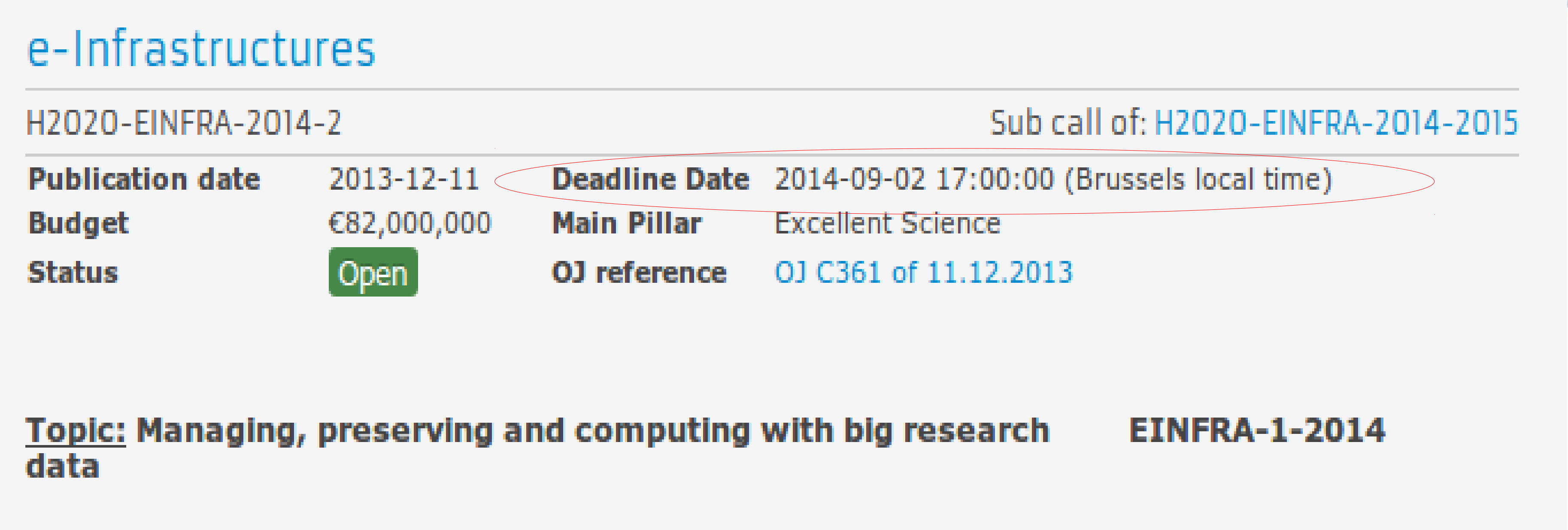
e-Infrastructures

H2020-EINFRA-2014-2 Sub call of: H2020-EINFRA-2014-2015

Publication date	2013-12-11	**Deadline Date**	2014-09-02 17:00:00 (Brussels local time)
Budget	€82,000,000	**Main Pillar**	Excellent Science
Status	Open	**OJ reference**	OJ C361 of 11.12.2013

Topic: Managing, preserving and computing with big research data **EINFRA-1-2014**

Specific challenge: Development and deployment of **integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructure**s incorporating advanced computing resources and software are essential in order to increase the **capacity to manage, store and analyse extremely large, heterogeneous and complex datasets**, including text mining of large corpora. These e-infrastructures need to provide **services cutting across a wide-range of scientific communities** and addressing a diversity of computational requirements, legal constraints and requirements, system and service architectures, formats, types, vocabularies and legacy practices of scientific communities that generate, analyse and use the data.



e-Infrastructures

H2020-EINFRA-2014-2 Sub call of: H2020-EINFRA-2014-2015

Publication date	2013-12-11	**Deadline Date**	2014-09-02 17:00:00 (Brussels local time)
Budget	€82,000,000	**Main Pillar**	Excellent Science
Status	Open	**OJ reference**	OJ C361 of 11.12.2013

Topic: Managing, preserving and computing with big research data **EINFRA-1-2014**

Specific challenge: Development and deployment of **integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructure**s incorporating advanced computing resources and software are essential in order to increase the **capacity to manage, store and analyse extremely large, heterogeneous and complex datasets**, including text mining of large corpora. These e-infrastructures

Aleph/CDF/LHCb importanti casi studio attorno a cui sviluppare un modello di LTDP per HEP e oltre --> altri esperimenti INFN hanno espresso interesse (AMS,AUGER,ARGO, KLOE,...) e la collaborazione con altri enti (INGV, INAF, …) è già stata avviata nella stesura di precedenti progetti (PIDES).

BACKUP

EINFRA-2-2014

Scope: Proposals should address all the following activities:

(1) Service-driven data e-infrastructure responding to general and specific requirements of researchers and research organisations for open access to and deposit of scientific information (including journal articles, books, monographs, conference proceedings, thesis, grey literature, software and data, as well as services linking literature, data and software). This e-infrastructure will further develop the research capacity through a coordinated and participatory architecture linking institutional and thematic repositories across Europe with scientific information to be used by humans and machines

[…]

Developing proof of concept and prototyping new services in support of open science (e.g. new forms of publishing, innovative services based on data mining, new forms of peer review etc.), assisting researchers and educators in everyday tasks.

[…]

(3) Supporting the global interoperability of open access data e-infrastructures and linking with similar initiatives across the globe in order to complement the physical access to research facilities with data access and to ensure that Europe plays a leading role in international collaborations.

It is expected that one proposal will be selected. A maximum of EUR 4 million of the total budget for this topic is foreseen for the article processing charges under point (2).

EINFRA-1-2014

(1) Establishing a federated pan-European data e-infrastructure to provide cost-effective and interoperable solutions for data management and long term preservation. The needs for data access, storage, replication, annotation, search, compute, analysis and reuse of information across disciplines should be accommodated in different research and education contexts. All these functions should expose standard interfaces for interoperation with other data sources to aggregate them or to be aggregated, considering also ethical and regulatory requirements for sensitive data (e.g. patient data). Sustainability is of paramount importance, therefore robust business models should be proposed to encourage investment from all stakeholders. Foreseen challenges are technical, legal and organisational, including engaging e-infrastructure operators and other service providers (such as those receiving support under topics EINFRA-2-2014, EINFRA-3-2014, and EINFRA-7-2014);

(2) Services to ensure the quality and reliability of the e-infrastructure, including certification mechanisms for repositories and certification services to test and benchmark capabilities in terms of resilience and service continuity of e-infrastructures;

(3) Federating institutional and, if possible, private data management and curation tools and services used across or at some point of the full data lifecycle, including approaches for identification of open data sources and data collected with sensitive or restricted access features. Services and tools should be federated on the basis of an open architecture and should offer or coordinate support to the development of Data Management Plans, in particular for Horizon 2020 project participants;

(4) Large scale virtualisation of data/compute centre resources to achieve on-demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers.

EINFRA-1-2014

(5) Development and adoption of a standards-based computing platform (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures…) to abstract application development and execution from available (possibly remote) computing systems. This platform should be capable of federating multiple commercial and/or public cloud resources or services and deliver Platform-as-a-Service (PaaS) adapted to the scientific community with a short learning curve. Adequate coordination and interoperability with existing e-infrastructures (including GÉANT, EGI, PRACE and others) is recommended

(6) Support to the evolution of EGI (European Grid Infrastructure) towards a flexible compute/data infrastructure capable of federating and enabling the sharing of resources of any kind (public or private, grid or cloud, etc.) in order to offer computing and storage services to the whole European scientific community. The proposal will address operations for supplying services (IaaS, PaaS, SaaS) at European level, engagement of and tailoring of services to new user communities and dissemination activities.

(7) Proof of concept and prototypes of data infrastructure-enabling software (e.g. for databases and data mining) for extremely large or highly heterogeneous data sets scaling to zetabytes and trillion of objects. Clean slate approaches to data management targeting 2020+ 'data factory' requirements of research communities and large scale facilities (e.g. ESFRI projects) are encouraged.

(8) Enable the creation of a platform and infrastructure for mining text aggregated from different sources/publishers that responds to the needs of users (researchers). This includes the definition of technical requirements (e.g. on interoperability, metadata standards and aggregation of new services) as well as addressing legal and contractual issues to serve the needs of text mining communities. The project should also provide consulting and counselling services to solve problems related with the legal framework and permissions to text mine collections, and to advise researchers on the benefits and practice of text mining. The development of the proposed platform and services should be informed by the studies on policy and licencing issues associated with Text and Data Mining that will be funded from the Call for “Developing governance for the advancement of Responsible Research and Innovation” in the "Science with and for Society" Work Programme (topic GARRI.3.2014 - Scientific Information in the Digital Age: Text and Data Mining). Therefore, the successful proposals in these two calls are expected to engage in a mutual dialogue and establish synergies in their work.