PEGASUS

M. Maggi
INFN Bari

Preserving and Ensuring an open Government for a smart Access to Scientific and cUltural Sources

coordinatore: Riccardo Pozzo (CNR/DSU)

PON R&C
2007-2013
regioni Convergenza
su
Long Term Digital
Preservation
per la ricerca
Open Access

Long Term Data Preservation nell'INFN

Storia

Le Osservazioni in AstroParticelle:

Meccanismi di open access per i dati raccolti da missioni scientifiche. Tradizione da NASA e ESA da cui derivano standard come FITS e protocolli di LTDP ISO come OAIS.

Le Misure in High Energy e High Density:

Solo recentemente la riproducibilità è insostenibile e i dati raccolti vengono immessi in procedure di LTDP con orizzonti comunque limitati.

Proposta Progetto Premiale PIDES

CNR/ITB, INAF, INFN e INGV vogliono sviluppare una piattaforma multidisciplinare per Long Term Data Preservation capace di archiviare l'informazione digitale ed il meccanismo di accesso e di utilizzo dei dati.

Applicazioni scientifiche prioritarie, di interesse degli enti di ricerca partecipanti, verranno usate per la raccolta dei requisiti e per l'adattamento dei sistemi agli eventuali standard già esistenti

Le applicazioni scientifiche in PIDES

- CNR/ITB: conservazione dei dati Omici e dei relativi sw (tecnologie genomiche, proteomiche, trascrittomiche and bioinformatiche)
- INAF: preservazione dei dati del centro di calcolo IA2 e dei codici di data ingestion, processamento e analisi con Standard dell'Osservatorio Virtuale
- INFN: preservazione dei dati dell'esperimento ALEPH, AMS, ARGO, AUGER-YGS, CDF, EAS-TOP, KLOE, MACRO, PAMELA
- INGV: raccolta e archiviazione dei sismogrammi storici

Attività definita in PIDES

Nell'INFN esistono già attività scientifiche a rischio di obsolescenza digitale

PIDES individua solo alcune da accogliere nelle infrastrutture del CNAF/Bologna

allo scopo di avere una infrastruttura pilota di data preservation

Il secondo sito è IA2/Trieste per l'INAF

PON su LTDP

Il PON è un'opportunità per includere, nella programmazione INFN di LTDP esperimenti che hanno grossi volumi di dati:

Gli esperimenti LHC

ALICE	←→	CATANIA
ATLAS	←→	NAPOLI
CMS	←→	BARI

Ovviamente con l'interesse di espandere ad altri domini il tema della LTDP e dell'Open Access

Infrastrutture

- INFN finanzia nei siti citati, Tier2 degli esperimenti che offrono capacità di calcolo e storage, limitati ad una utenza limitata.
- Attraverso la GRID l'INFN ospita il calcolo di altri domini scientifici

- RECAS 

RECAS

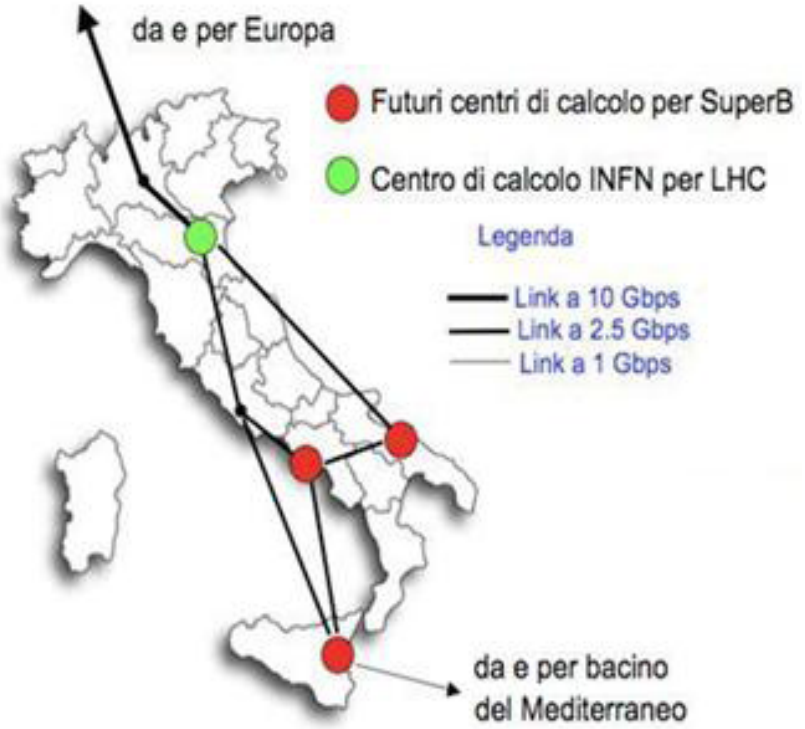
- INFN

(Bari, Catania, Cosenza, Napoli)

- UNIBA
- UNINA

1 core = 8-12 HepSpec06

	Potenza elaborative (kHepSpec)	Storage (PByte)
UNINA	6	0,8
INFN-NA	2	0,3
UNIBA	10	2,5
INFN-BA	3	0,5
INFN-CT	7	0,8
INFN-CS	5	0,6
TOTALI	33	5,5



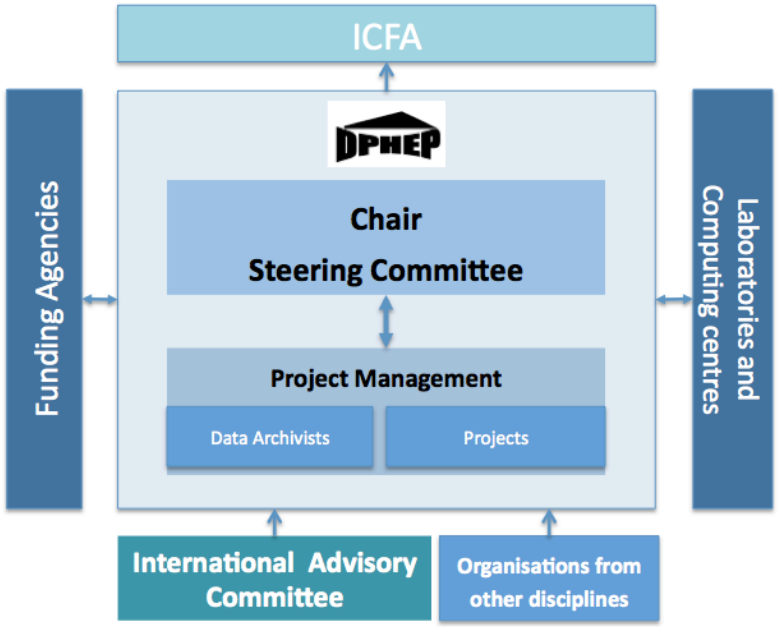
Logistica in grado di ospitare ulteriori risorse infrastrutturali

DPHEP

- Pannello dell'ICFA dal 2009

https://dl.dropbox.com/u/48384809/dataprese/DPHEP_2012-05.pdf

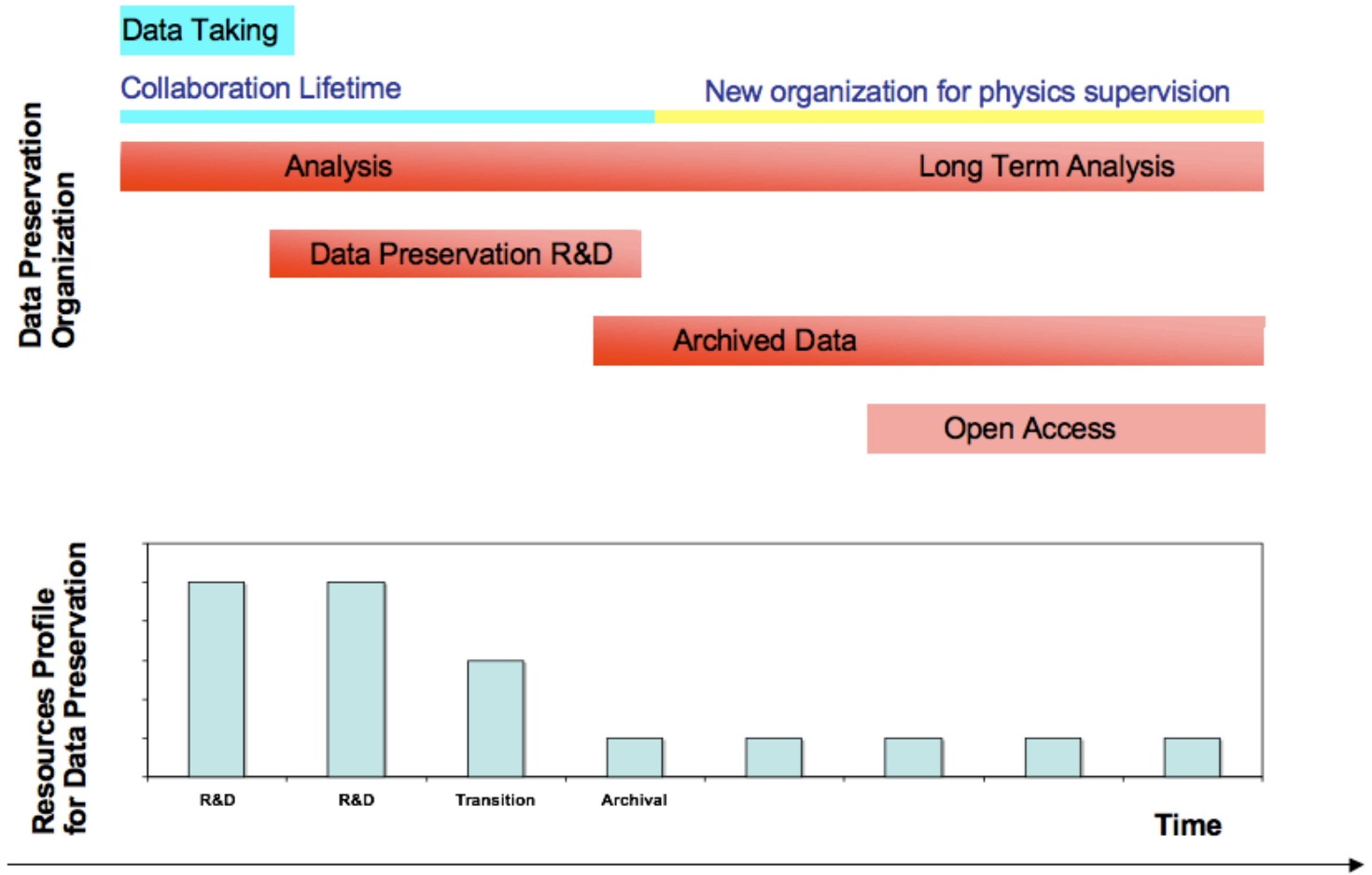
- Partenza di una collaborazione internazionale



Incentivare collegamento con altre organizzazioni attive in questo campo

HORIZON 2020

Experiment Timeline



DPHEP Models

Preservation Model	Use Case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analysis
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

OAIS

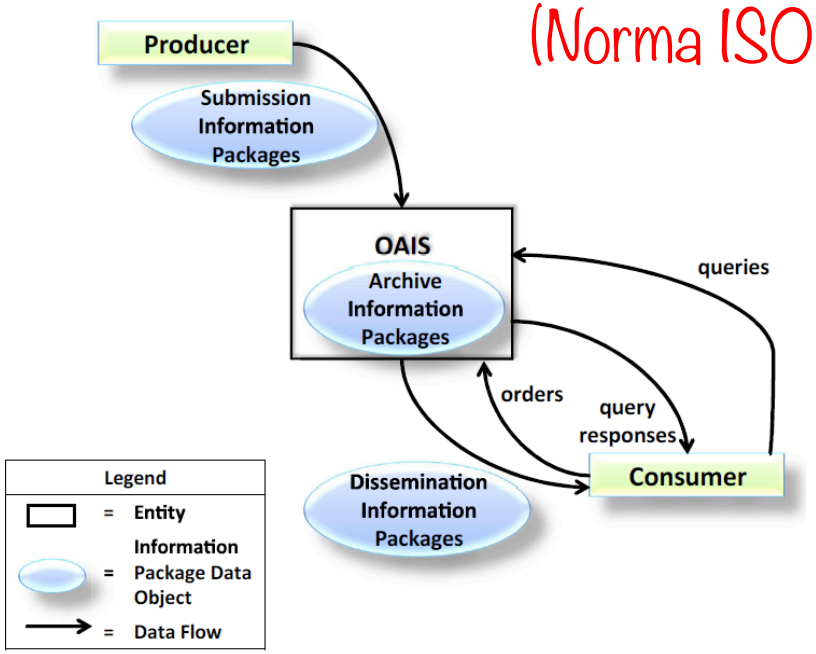
Consultative Committee for Space Systems

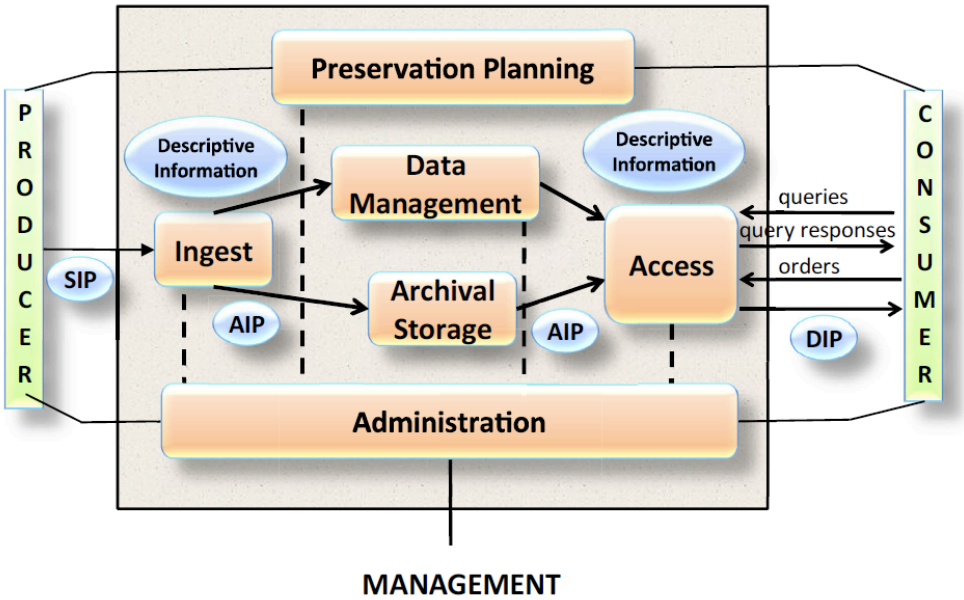
Recommendation for Space Data System Standards

REFERENCE MODEL FOR AN

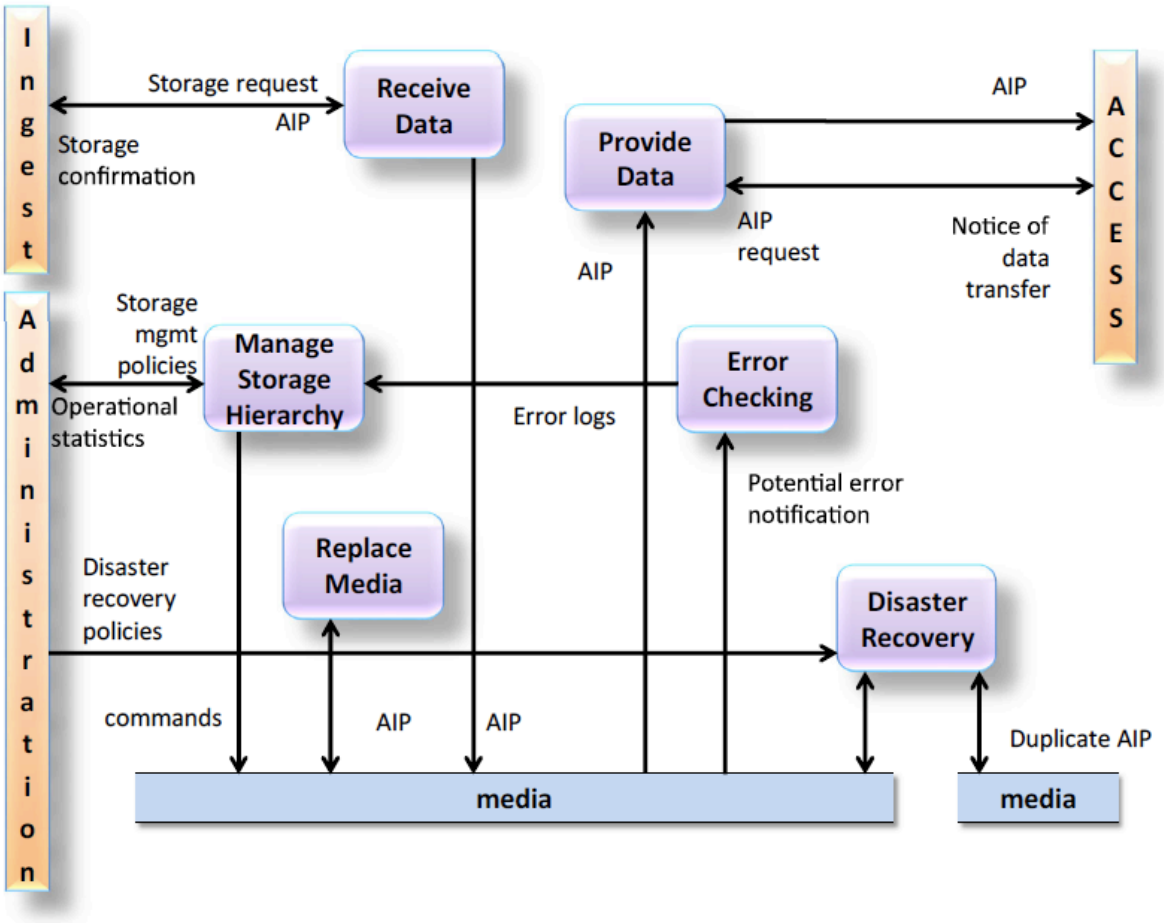
Open Archival Information System

(Norma ISO 14721:2003)





OAIS Archival Storage



Progetti Internazionali 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

Progetti "Regionali"

DASPOS (USA)

(Data And Software Preservation for Open Science)

- Multi-disciplinary effort recently funded by NSF
- **Discovery & Coordination:** Several Workshop to be organized
- **Prototyping & Experimental Task:** Create Data Model & Query Semantics Define Elements of Software Reproducibility

PREDON (FRANCIA)

(PRÉservation des DONnées)

- Demonstrate The Interests Of Several national Labs In Complex Scientific Data preservation
- Communication: Exchanges, Workshop and White book specifications
- Demonstrator of multi-discipline access and preservation unit (focus on scientific complex data)
- Install "National Data Observatory"

PON per LHC

- Rafforzando la infrastruttura ReCaS è possibile avere una parte della infrastruttura elettronica dedicata alla LTDP per LHC.
- Ogni sito deve svolgere attività per qualunque esperimento
- Il grosso delle risorse sono di storage.
  - Livello 1: documentale, poco spazio ma molto performante
  - Livello 2: Disco standard (~4 PB)
  - Livello 3-4: Tape (10 PB)

Necessario un link veloce per permettere l'utilizzo delle risorse distribuite e non taggate su un esperimento

Sinergie

- SCIDIP-ES rilascerà le prime release per LDTP basato su OAIS nel 2013
- INSPIRE.NET per la parte documentale OPEN ACCESS che usa INVENIO digital library tech
- RECAS per la logistica
- EUDAT per il bit preservation
- PIDES se approvato
- DPHEP e LHC per il sistema di LTDP da ospitare

PON oltre a LHC

- L'INFN è interessato ad avere obiettivi scientifici/ umanistici più ampi (Vedi PIDES, DCH-RP, etc.) ospitando i relativi programmi di LTDP
- Si riconosce l'importanza dell'Open Access. Riferimenti ai progetti internazionali esistenti OpenAIRE, OpenDOAR, Datacite, ENGAGE

Proposta Progettuale PEGASUS

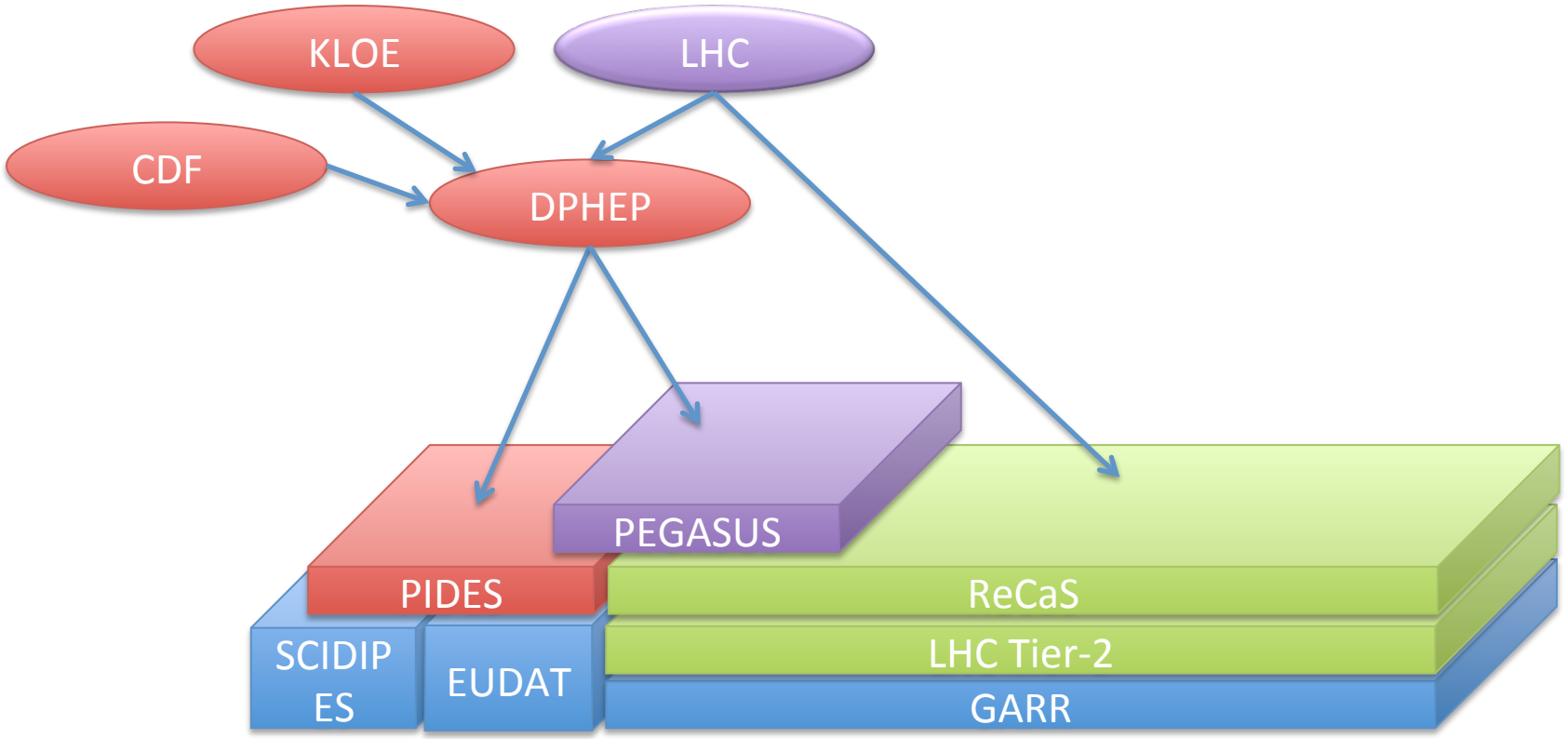
Piattaforma di Digital Library

che poggia su

Infrastruttura di Storage Distribuita

e che utilizza alcune periferiche per l'acquisizione di record digitali per i beni culturali

Le linee scientifiche hanno già i loro record digitali


KLOE
LHC
CDF
DPHEP
PEGASUS
PIDES
ReCaS
SCIDIP
ES
EUDAT
LHC Tier-2
GARR

Info Progetto

- CNR (Ba-Cs-Ct-Na)
- ENEA (Portici)
- INFN (Ba-Ct-Na)
- INAF (Ct)
- CINECA (Cs)
- UNIBA
- UNICAL
- UNICT
- UNINA

Presentazione proposta:
3-Apr-2013

Budget: 10 M€

Completamento Attività:
30-Mar-2015

Rendicontazione Finale:
30-Giu-2015

Spese Ammissibili:
Ferro...

Status Progetto

- Attesa risposta inizio Giugno

Altre proposte:

1) Medical Images Database and visual memory engine. CEINGE, Biot. Avanzate S.C., UNINA, UNICatanzaro, Ist.Naz.Tumori
2) Costituzione di una infrastruttura per la raccolta e condivisione dei risultati della ricerca e attivazione di percorsi clinici integrati. Ist. Centro Neurolesi Bonino Pulejo, UniMediterranea, UNISA
3) Digital Saving System finalizzato all'armonizzazione di una Rete di Biobanche delle regioni di Convergenza. Fond. Cutino Onlus, Ist. Tumori Giov.Paolo II, Fond RiMed, UNIPA
4) Informazione multimediali per Oggetti Territoriali. UNINA, UNI OrsolaBenincasa, UNI Parthenope, POLIBA, UNINAII, UNIPA, UNISA

Budget Splitting -1

- Calabria: 2,695,000 €
- Campania: 2,517,260 €
- Puglia: 2,108,840 €
- Sicilia: 2,678,900 €

Potenziamento: 9,145,750 €

Formazione: 854,250 €

Budget Splitting -2

- CNR 1,903,910 € (comprensivi costi progetto)
- INFN 1,850,000 €

- UNIBA 1,016,840 €
- CINECA 1,000,000 €
- ENEA 950,000 €
- INAF 770,000 €
- UNICAL 725,000 €
- UNINA 465,000 €
- UNICT 465,000 €

HEP Physics: 2,880,000 €

per LTDP LHC storage

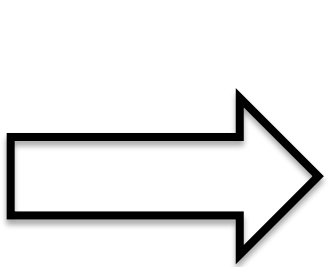
1) Attività iniziate e da farsi riconoscere da exp
2) Risorse pledge-abili
3) Liberare risorse al CNAF

Attività I

Digital Library:

- CINECA ha soluzione fatta con La Sapienza
- CNR ha progetto MIUR finanziato per Biblioteca Digitale per la Scienza e Tecnologia
- HEP ha INVENIO su cui si basa INSPIRE, e che integra HepData e Rivet

Ingestion
Processing
Dissemination
Curation



Digital
Preservation

Attività 2

HEP LTDP

Creazione di un'infrastruttura Tier1-like

- Tape Library con > 12 PB (Ba)
- Disk Storage 5 PB (Ct-Na)
- + Richieste "Ancillari"

Fondamentale la connessione (O(100) Gbps) tra i siti e verso il CNAF per operare il sistema centralmente

Logistica ReCaS già finanziata e in avvio

Implicazioni

- Se approvato PEGASUS offre l'opportunità all'ente di fornirsi di un'infrastruttura che gli esperimenti vogliono usare
- Il progetto premiale PIDES prevedeva esperimenti di Gruppo II, Tevatron e LEP, quindi PEGASUS punta su LHC
- Questo pone un problema di gestione per un sistema di storage e calcolo "veramente" distribuito