Preservazione dell'Informazione Digitale ed Eredità Scientifica

Marcello Maggi
INFN Bari

Definizione

Data Preservation:

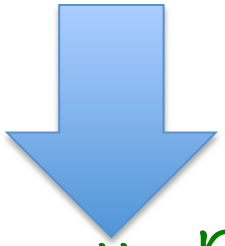
Meccanismo di conservazione della capacità di utilizzare dati digitali molto oltre il limite di esistenza dell'organizzazione che ha prodotto tali dati ovviando all'obsolescenza digitale

Include il Bit Preservation (ma molto altro)

Esclude la pronta disponibilità

Coordinamento degli EdR

Grazie all'Italian Grid Infrastructure (IGI) è stato costruito un coordinamento multi-disciplinare tra CNR, INAF, INFN e INGV



Proposta di un Progetto Premiale MIUR per Long Term Data Preservation

PIDES

CNR, 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

- Obiettivi
- Quadro dei progetti esistenti
- Fasi realizzative
- Strategia verso HORIZON 2020

Le applicazioni scientifiche

- CNR: 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 CDF
- INGV: raccolta e archiviazione sei sismogrammi storici

4 casi concreti prioritari

Obiettivi

1) Definire e realizzare una piattaforma per la preservazione a lungo termine dei dati digitali di grandi dimensioni prodotti nell'ambito di applicazioni scientifiche di interesse degli Enti partecipanti.

2) Determinare le strategie per l'accesso e l'utilizzo dei dati per il periodo di conservazione. Sviluppare soluzioni che permettano di preservare l'usabilità delle applicazioni per l'elaborazione e l'analisi dei dati.

3) Definire formati e implementare protocolli standard per la preservazione dei dati che eliminino o riducano, almeno per una parte di essi, le dipendenze da applicazioni o domini scientifici specifici e ne permettano la pubblicazione. Stabilire le procedure di validazione e i sistemi di monitoraggio per garantire l'effettiva utilizzabilità dei dati preservati.

4) Definire meccanismi per la realizzazione di sistemi integrati di preservazione dei dati digitali distribuiti in rete.

5) Adattare gli standard già esistenti in alcuni domini specifici alla nuova piattaforma per la preservazione dei dati digitali.

6) Identificare e affrontare le problematiche di integrità, sicurezza e riservatezza dei dati che rappresentano un aspetto critico per l'affidamento di dati sensibili o di elevato valore economico.

7) Identificare le opportunità per Education, Training and Outreach

Le attività di Definizione e Sviluppo
vanno intese in un contesto multidisciplinare e internazionale

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

OAIS

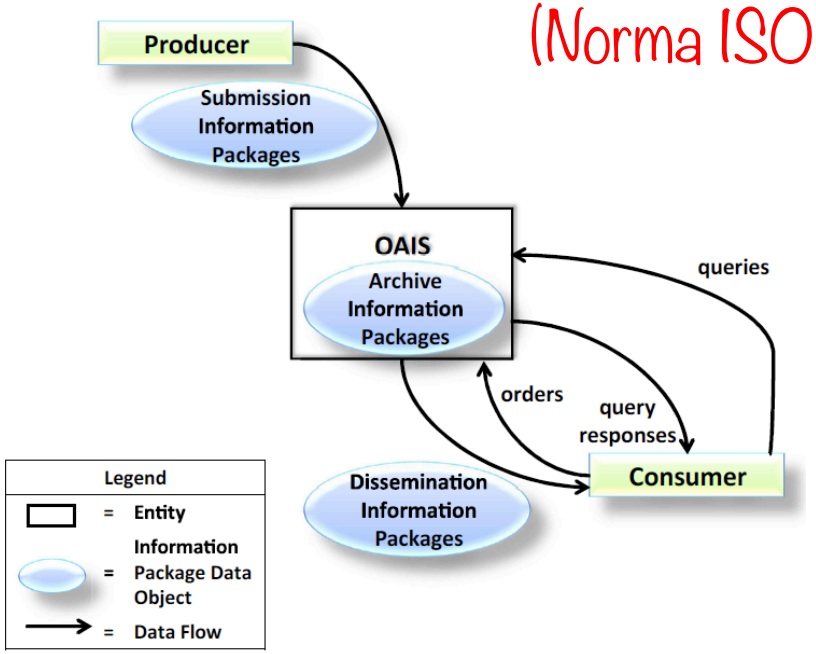
Consultative Committee for Space Systems

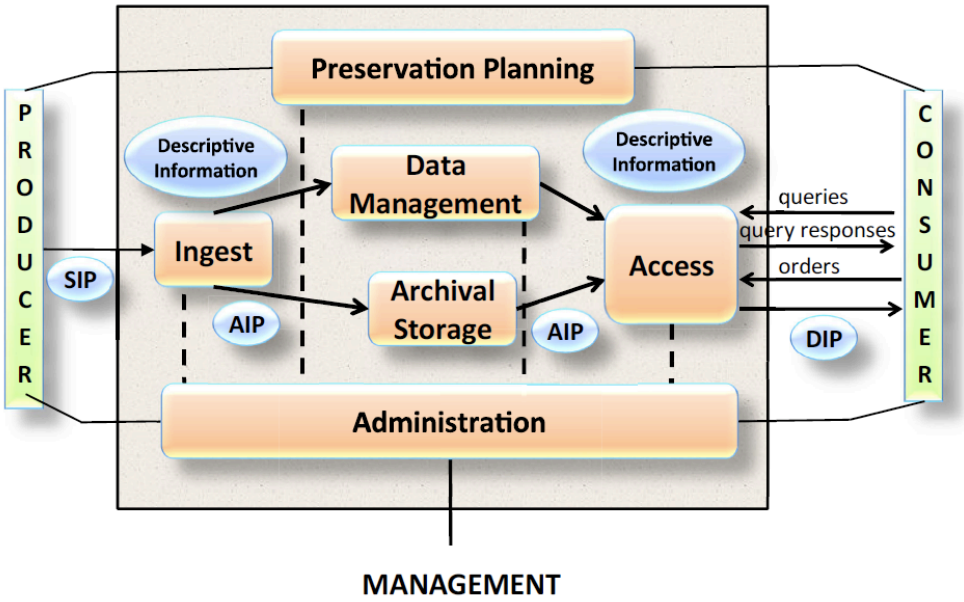
Recommendation for Space Data System Standards

REFERENCE MODEL FOR AN

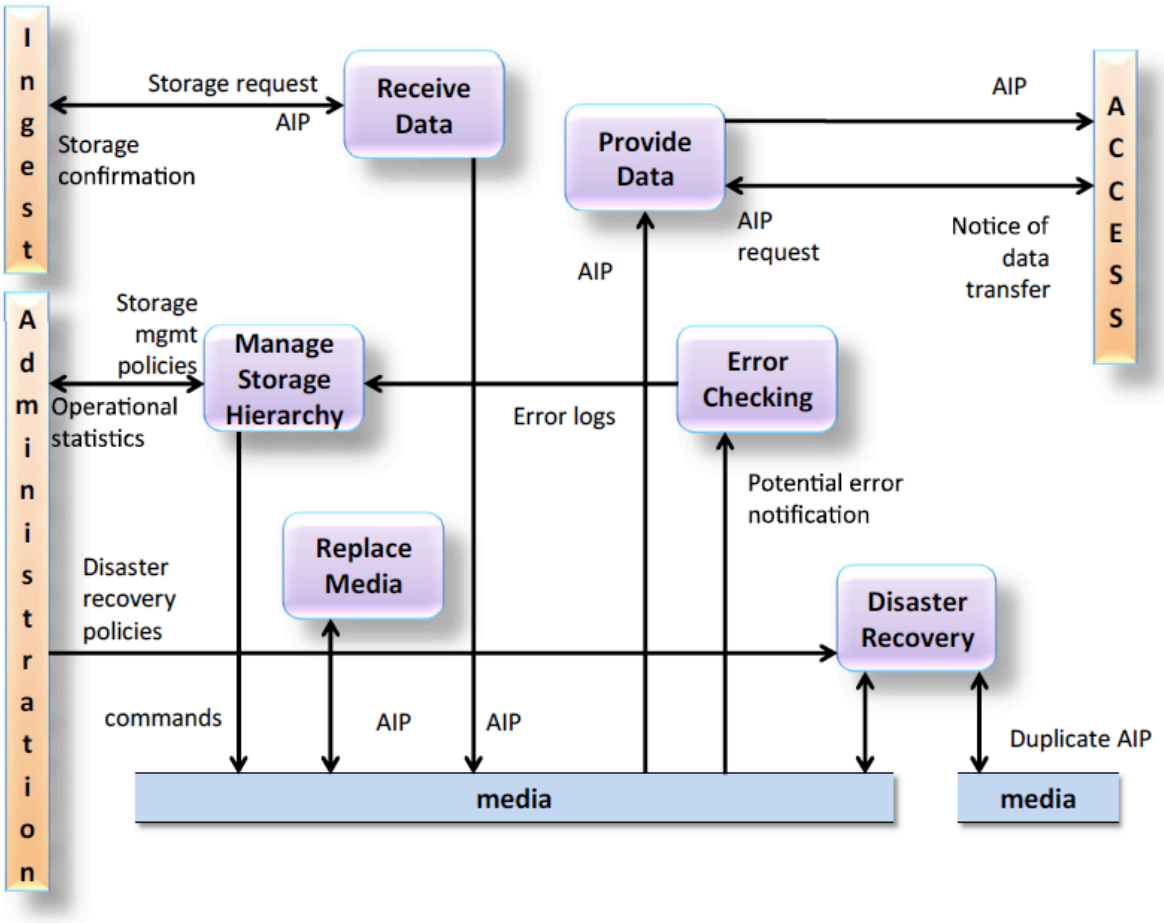
Open Archival Information System

(Norma ISO 14721:2003)





OAIS Archival Storage

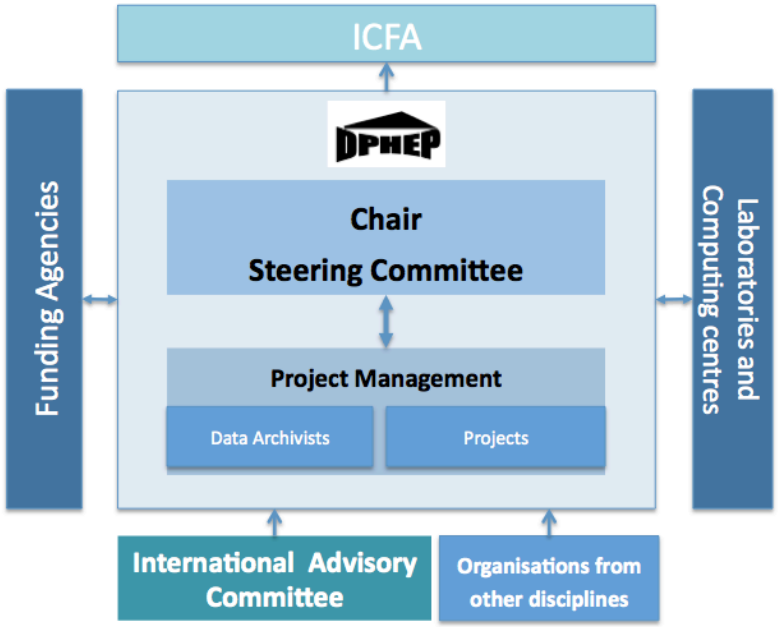


DPHEP

- Prodotto documento finale per ICFA

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

- Partenza di un progetto finanziato



Incentivare collegamento con altre organizzazioni attive in questo campo

HORIZON 2020

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

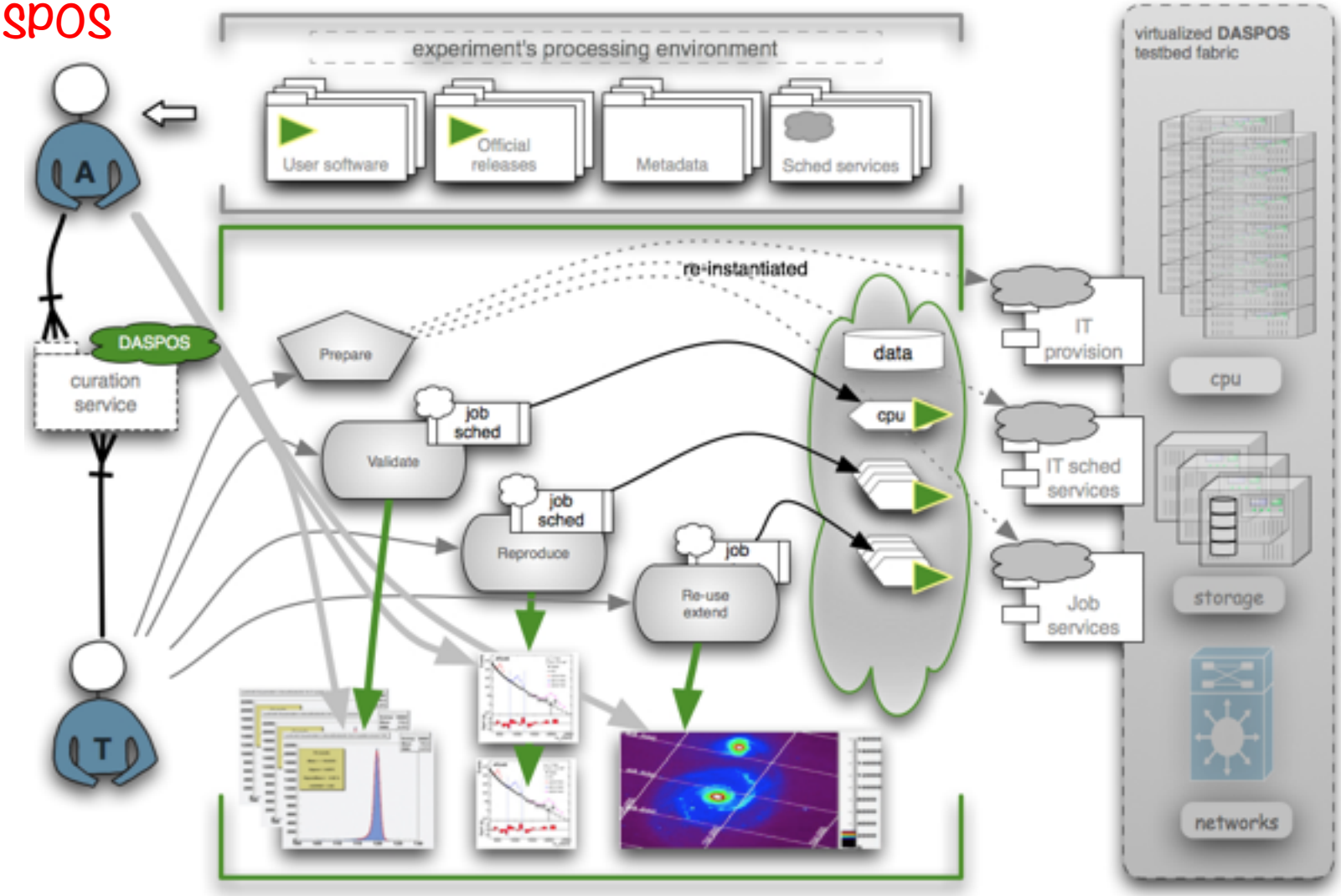
PREDON (FRANCIA)

(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"

Curation Challenge

DASPOS



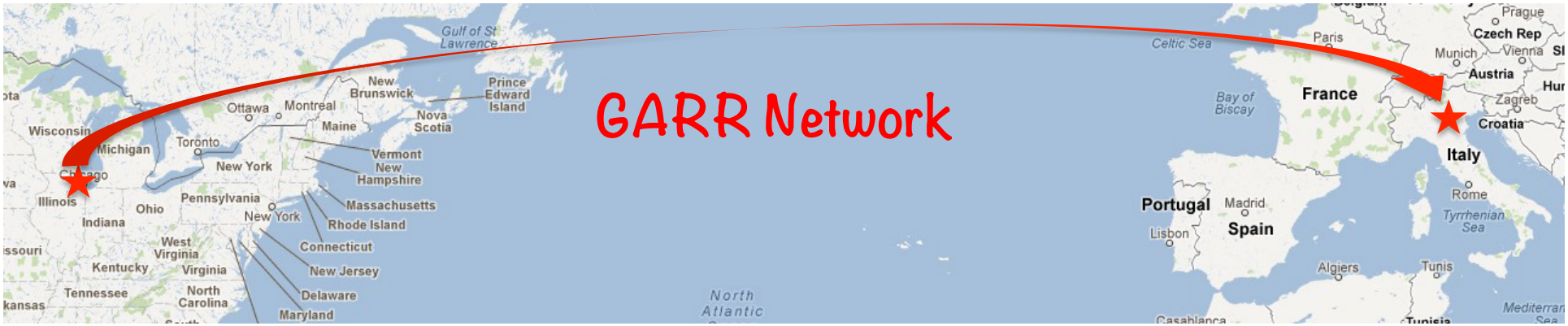
Fasi di realizzazione

Fase 1) Raccolta delle esigenze nelle diverse comunità scientifiche. Definizione dei protocolli e dell'architettura per l'accesso dati e l'usabilità delle applicazioni. Realizzazione di un primo prototipo di sistema di archivio di almeno 4 PB

Fase 2) Costruzione di un prototipo di un Repository Digitale estensione del sistema di storage a dimensioni opportune (~10 PB) e nella implementazione dei servizi fondamentali quali Trust, Accounting, Integrity, Redundance, Fault Tolerance, Identity management. Finalizzazione dell'accesso ai dati con tecnologie di virtualizzazione. Inizio dell'attività di Dissemination.

Fase 3) Collaudo e messa in opera di un sistema distribuito di Repository Digitale in grado di garantire i servizi di Ridondanza, le procedure di Disaster Recovery, la messa in produzione dei Controlli di accesso, i sistemi di Sicurezza e Confidenzialità. Continuazione della Dissemination, Training per l'uso dell'infrastruttura di preservazione digitale e costruire il sistema di Outreach per le scuole e l'università

Il punto di inizio



Data (~10 PB) → Stored on Tape (LT03, LT04, T10K) → Three formats:
1) Raw
2) Production
3) Ntuples

Preservazione dei dati e del sw dell'esperimento
Data Access System
Condition Data and Metadata

Work Packages

Work Package	Title	Description	Funding Agencies
WP1	Management	Coordination	CNR, INAF, INFN and INGV
WP2	Domain Specificity	Use cases for the different communities are used to gather requirements and to adapt existing data preservation systems to the specific needs	CNR, INAF, INFN and INGV
WP3	Architecture	Implementation and extension of data preservation standards like OAIS[*]	CNR,INAF, INFN and INGV
WP4	Bit Preservation	Storage System definition, control system, integrity, redundancy and disaster recovery implementation	INAF and INFN

[*]
"Reference Model for an Open Archival Information System (OAIS).
CCSDS 650.0-B-1, Blue Book, January 2002"
https://public.ccsds.org/publications/archive/650x0b1.pdf;
Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August 2009.

Work Packages

Work Package	Title	Description	Funding Agencies
WP5	Code Preservation	Development of data access system to preserve scientific applications through virtualization techniques	CNR, INAF and INFN
WP6	TestBed & Validation	Deploy and Standardize Validation procedures to ensure access and usability of the data	CNR, INAF, INFN and INGV
WP7	Data Access Policies	Trust, Security, Intellectual Property, Encription, etc.	CNR, INAF, INFN and INGV
WP8	Dissemination, Training & Outreach	Documentation, Event Organization for the dissemination and consolidation of data preservation. Organization of training session for digital repository administrator. Activation of a framework for scholastic use of scientific data	CNR, INAF, INFN and INGV

HORIZON 2020

Long Term Data Preservation

Data preservation può essere un elemento chiave nelle tre priorità di HORIZON 2020

- **Excellent Science:** Long Term Data Preservation può essere basato su una e-infrastructure che necessita di essere sviluppata ed operata in connessione con le infrastrutture sceintifiche esistenti.
- **Industrial Leadership:** Tra le attività ICT "Content technologies and information management" è un tema dove la preservazione dei dati può trovare il suo spazio ed un'opportunità per collaborare con partner industriali
- **Social Challenges:** La Commissione ha l'obiettivo di rafforzare le basi scientifiche e tecnologiche attraverso la European Research Area (ERA). La preservazione dei dati è un'attività di supporto a ERA. L'accesso a dati scientifici per Education e Outreach saranno temi chiavi nella linea "Innovative Societies".

Conclusioni

- La proposta sarà rivista alla luce del bando dei premiali 2012
- Vuole essere parte di uno sviluppo che vuole trovare soluzioni concrete alla data preservation in casi specifici prioritari
- Vuole nascere da un coordinamento multidisciplinare "Regionale" che intende essere parte du una collaborazione "Globale" per Horizon 2020