Data Preservation and Scientific Heritage in the Italian Community

Marcello Maggi
INFN Bari

Coordination of Italian Agencies

Thanks to Italian Grid Infrastructure (IGI) we have build a multi-disciplinary coordination within

CNR, INAF, INFN and INGV

The first step is the proposal of a project to obtain funding from Italian Minister of Research (MIUR), devoted to multi-disciplinary issues, for

Long Term Data Preservation

PIDES Outlook

CNR, INAF, INFN e INGV intend to realize a multi-disciplinary platform for Long Term Data Preservation able to archive the digital information and the mechanism of data access and data analysis.

Typical use case of the different scientific communities will be used to collect the requirements and to adapt existing data preservation systems to eventually existing standards.

- Objectives
- Framework within existing projects
- Realization phases
- Strategy in view of HORIZON 2020

The Use Cases

- CNR: omics-data and relative analysis sw tools (data from genomics, proteomics, trascrittomics and bioinformatics tools)
- INAF: data from IA2 centre with data ingestion sw tools
- INFN: CDF data preservation
- INGV: preservation of the historical seismograms

4 concrete cases considered as priorities

Objectives

1) Define and Realize a Long Term Data Preservation Platform hosting Scientific Applications
2) Determine the data access strategies during the preservation period and develop solutions that allow the data analysis
3) Define formats and implement standard protocols in data preservation to avoid or mitigate the dependencies to external libraries and to allow open access. Establish validation procedure and monitoring system to ensure the correct data usability
4) Define the mechanism to build integrated system for preservation of distributed data
5) Adapt existing standards in some specific domain within the Platform
6) Identify and face critical issues as integrity, security and privacy for sensible data
7) Identify Education, Training and Outreach opportunities

The Definition, Implementation, Development, etc. should be intended in a international and multi-disciplinary contest

Existing Projects

SCIDIP-ES

(SCIence Data Infrastructure for Preservation - Earth Science)

- Main Long Term Data Preservation projects. 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.
- This project intends to start realizing OAIS components

EUDAT

(EUropean DATa infrastructure)

- Project to build an e-Infrastructure where data can be exchanged and shared under well defined services and standards.
- It has data preservation topic, but as bit preservation and something more

Project Realization

Phase 1) Gathering the needs of the different Scientific communities. Definition of the protocols and of the architecture for data access and scientific applications. Realize a prototype for a 4 PB archive system

Phase 2) Build a prototype of a ≈10 PB Digital Repository, which implements Trust, Accounting, Integrity, Redundancy, Fault Tolerance, Identity Management. Finalize data access with virtualization mechanisms. Start Dissemination

Phase 3) Commissioning and Deployment of a distributed Digital Repository system, which realize redundancy and disaster recovery procedures, access control security and privacy. Organization of the training for the use the digital preservation platform and build the Outreach and Education system

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

[*] Norma ISO 14721:2003;
"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.

OAIS

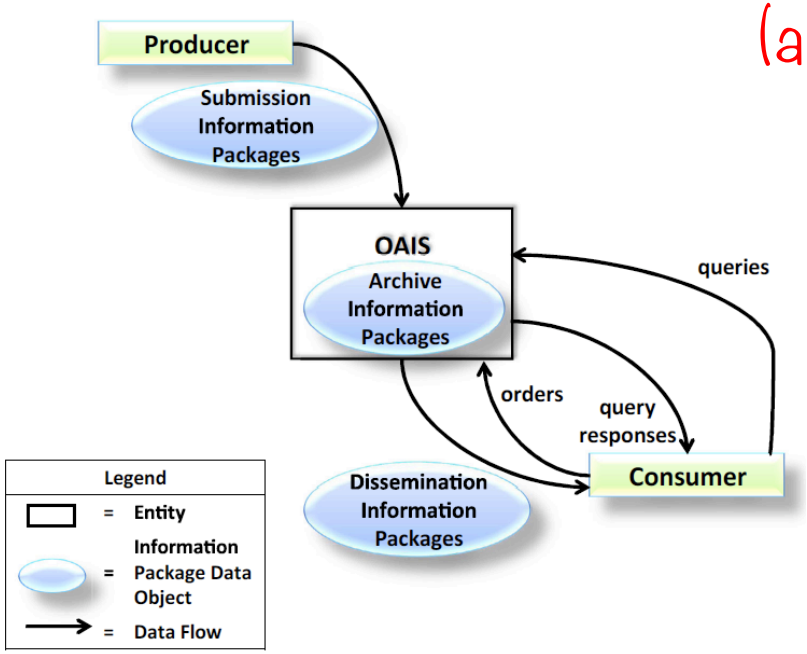
Consultative Committee for Space Systems

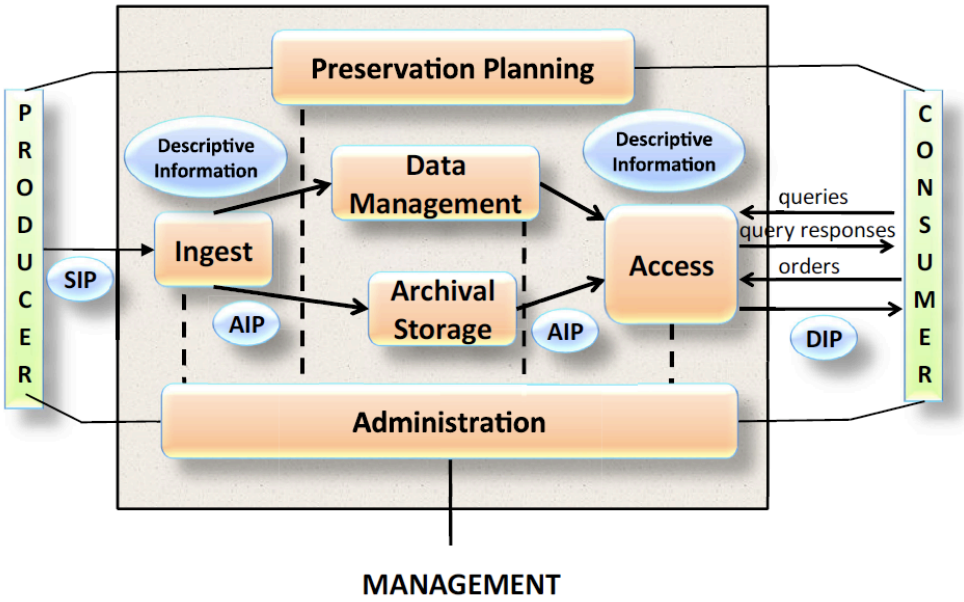
Recommendation for Space Data System Standards

REFERENCE MODEL FOR AN

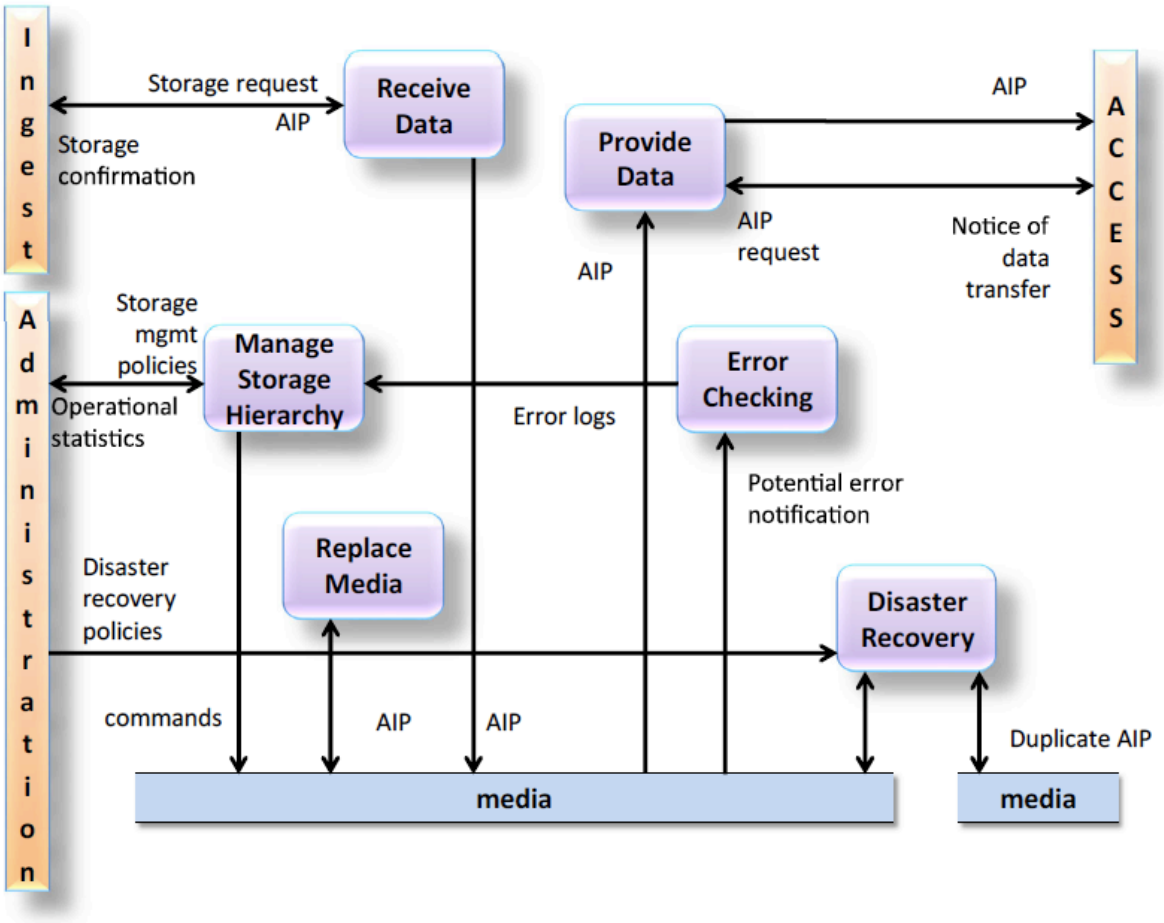
Open Archival Information System

(an ISO)





OAIS Archival Storage



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

Funding Long Term Data Preservation

- Data preservation can be a key issue in the three priorities of HORIZON 2020

- **Excellent Science:** Long Term Data Preservation should be based on a e-infrastructure that need to be operated in connection with existing scientific infrastructure
- **Industrial Leadership:** among the ICT activities, "Content technologies and information management" is a topic where data preservation can find industrial partnership and create public-private network
- **Social Challenges:** One of the objectives of Eu Commission is to enforce the scientific and technological basis through the European Research Area (ERA). Data Preservation is a support activity in ERA. Scientific Data Access for Education and Outreach will be key issues in "Innovative Societies"

Final Remarks

- This proposal will be reviewed and more news should be available within 2012
- It intends to have concrete useful preservation solutions, and maximize the the exchange experiences
- It is a "Local" multi-disciplinary network, that would like to be part of a "Global" network in view of Horizon 2020