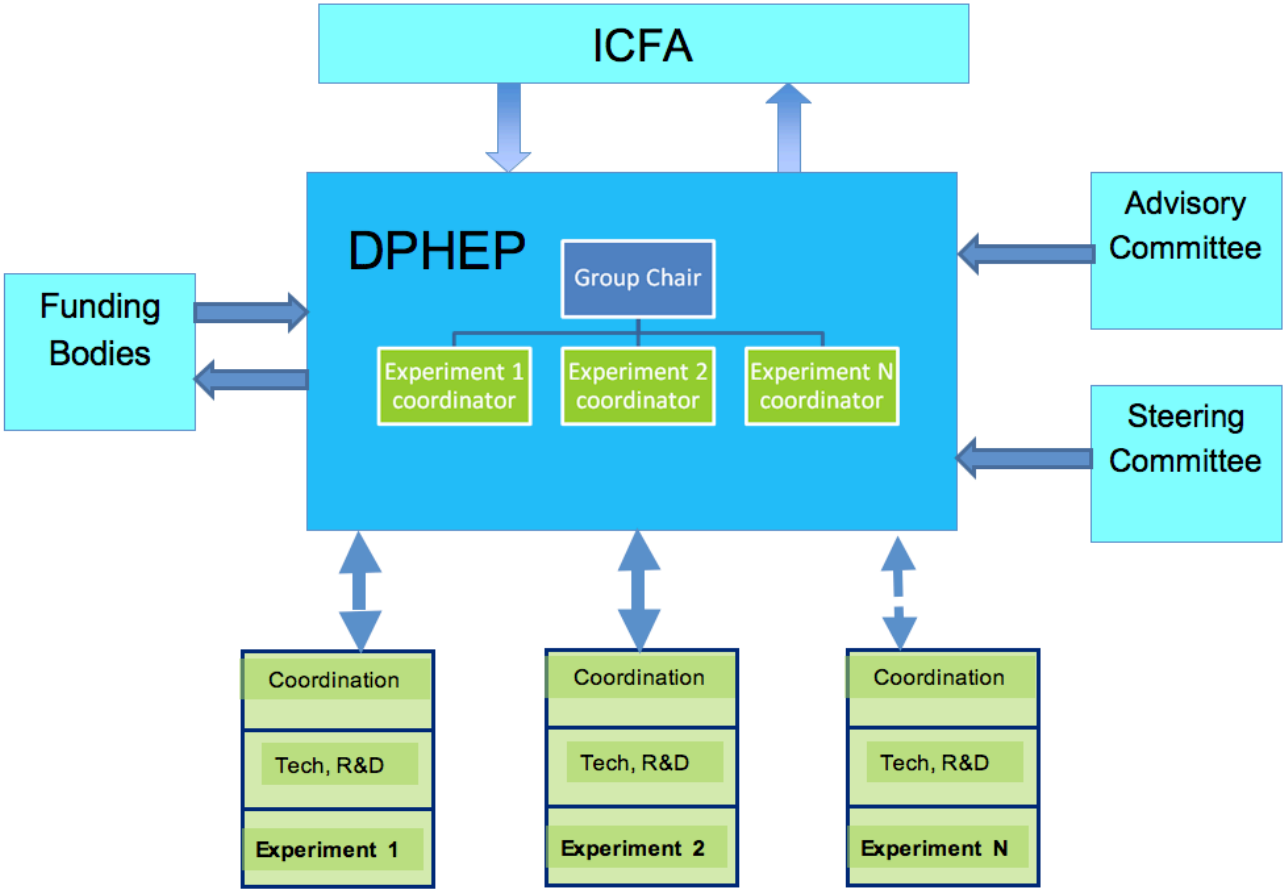
Data Preservation in HEP
strategie per HORIZON 2020

M. Maggi
INFN-Bari

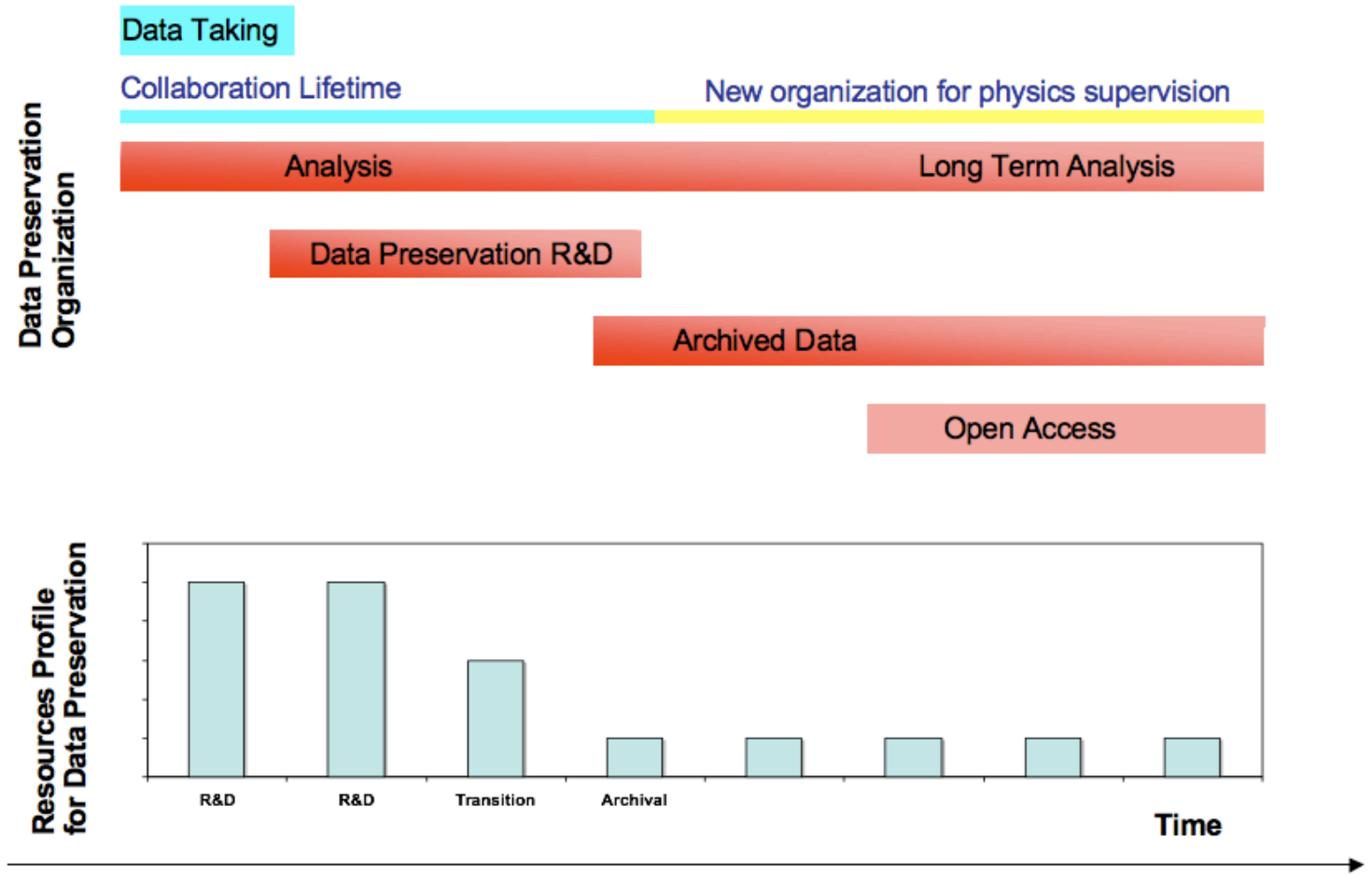
Summary

- Data Preservation in HEP
- PODDS proposal: Support Action in FP7
  (goal: build a DP project within HORIZON 2020)

DPHEP study group



DPHEP timeline



DPHEP deliverables

Objective	Deliverable
1.Positioning as forum	Catalogue of technical knowledge and practical solutions Description of possible alternatives for governance
2.Co-ordination of projects	Common R&D projects meet the expectations of the stakeholders
3.Harmonisation and liaison	Synchronisation of preservation projects in the field. Addition of external knowledge to HEP
4.Design sustainable future	Characterization of discipline-wide toolkit for preservation Business plan for long-term preservation in HEP
5.Outreach and advocacy	Understanding of needs/opportunities for medium- and small-sized collaborations Presentations to and agreements with funding agencies.

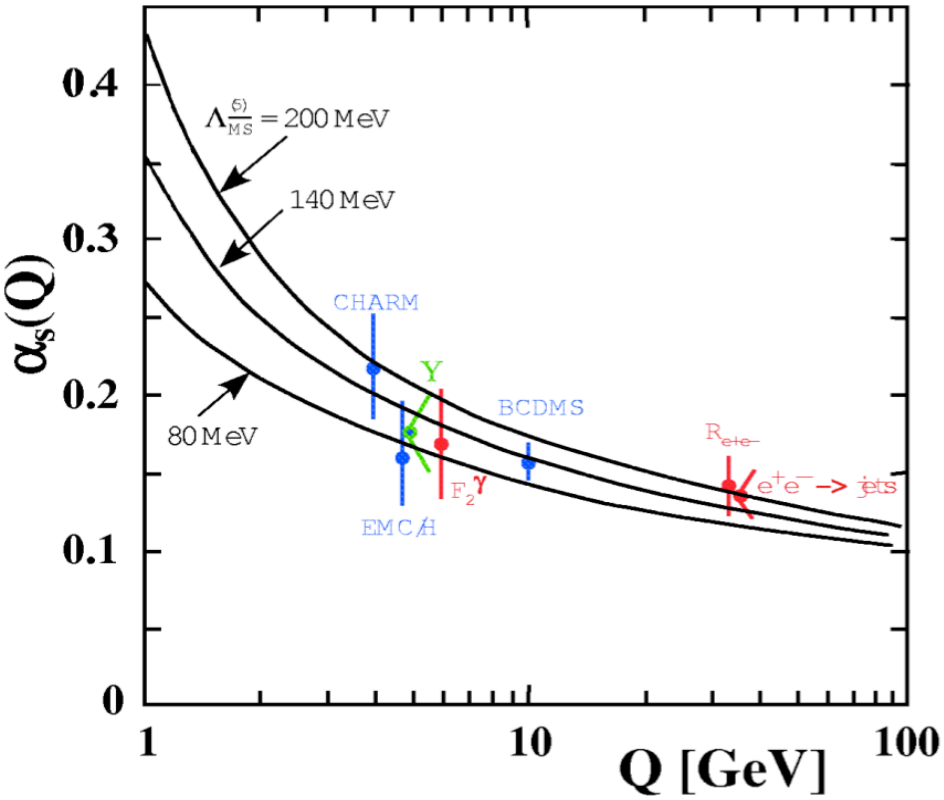
Table 6: Deliverables of the DPHEP Study Group.

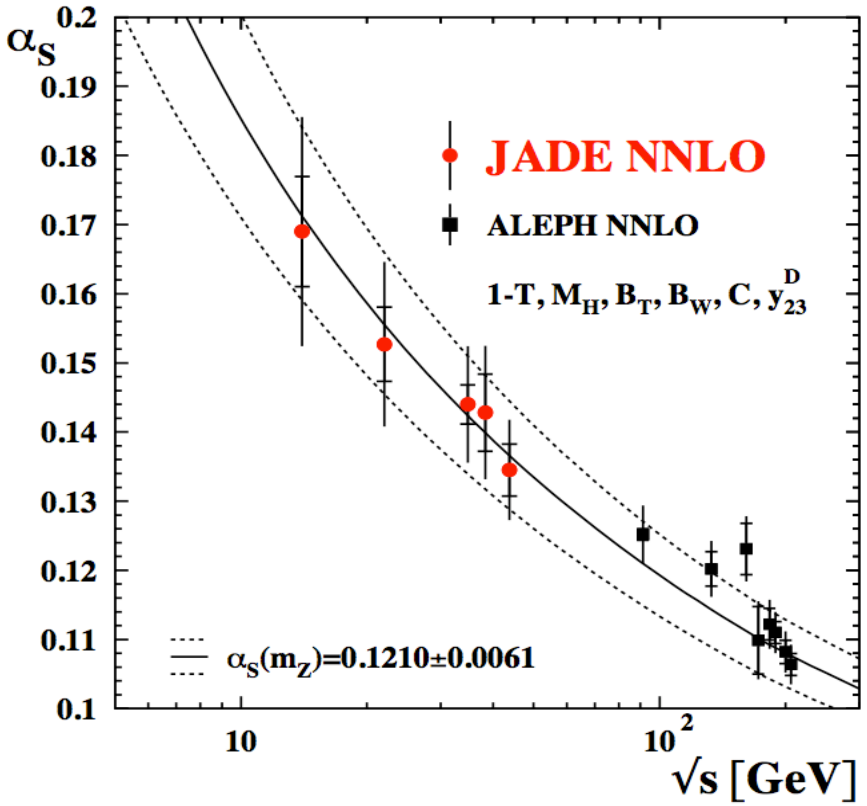
DPHEP models

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
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

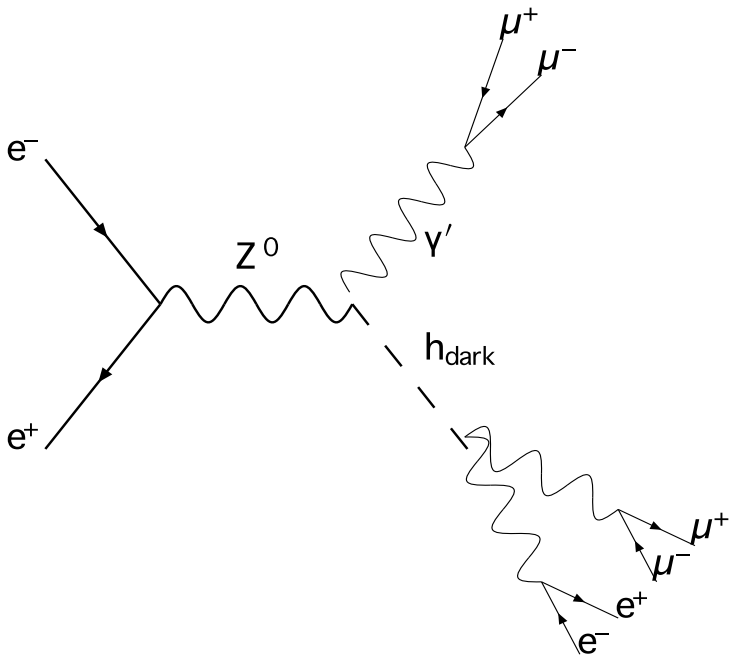
Table 3: Various preservation models, listed in order of increasing complexity.

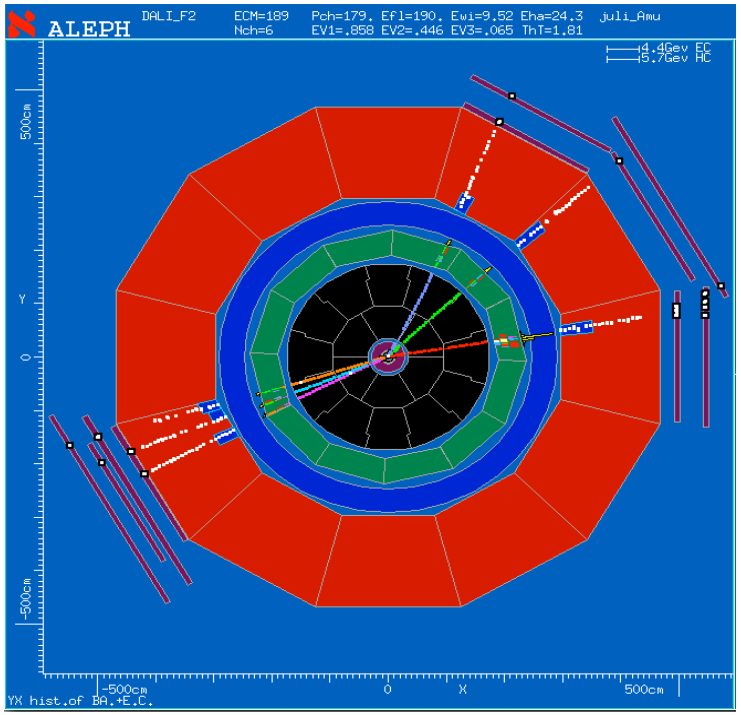
Importance of Shared Data





Need to preserve data production





The entire ALEPH Monte Carlo data production chain had to be used

Increasing Interest in HEP

- CDF
- D0
- BaBar
- Belle
- Zeus
- H1
- BES III
- Existing supporting documents from LHC

FP7 ICT Call 9

ICT Work Programme 2011-2012 Objective 4.3

Digital Preservation

Objective

creating technology solutions and innovative methods

for keeping **digital resources available and useable over time**

EU Investment

since 2006: 15 projects; € 86 mio EU-funding

PODDS project

Preservation Of Data for the Digital Society

Support Action

CERN/INFN/MPG

~700 kEuro/2years requested

to

- underpin the implementation of the programme
- help in **preparations for future EU research activities** **Seed Project for HORIZON 2020**

PODDS purpose

- To consolidate and disseminate through **well-established channels the current state of the art in terms of research into digital preservation – with a strong focus on reuse – as well as existing techniques and known challenges;**
- To use previous, current and planned projects, networks and fora to further **promote the fundamental necessity of providing solutions to a documented and agreed set of use cases;**
- To develop a multi-disciplinary medium- to long-term **roadmap for the digital preservation domain, clearly identifying the steps necessary to establish pilot services in this area and mechanisms whereby they could be validated from a technical and sustainability viewpoint, leading to a pan-European / truly international set of service offerings in the coming five to ten years. This roadmap should be endorsed as widely as possible by recognized institutes, bodies and communities.**

PODDS strategy -1

The consolidation of existing requirements, knowledge and research results into a single knowledge repository that is as broad as possible in terms of the disciplines that it represents. This includes not only sciences (High Energy Physics, Astronomy and Astrophysics, Life and Earth Sciences and so forth), but also arts and humanities as well as industry and service sectors;

PODDS strategy -2

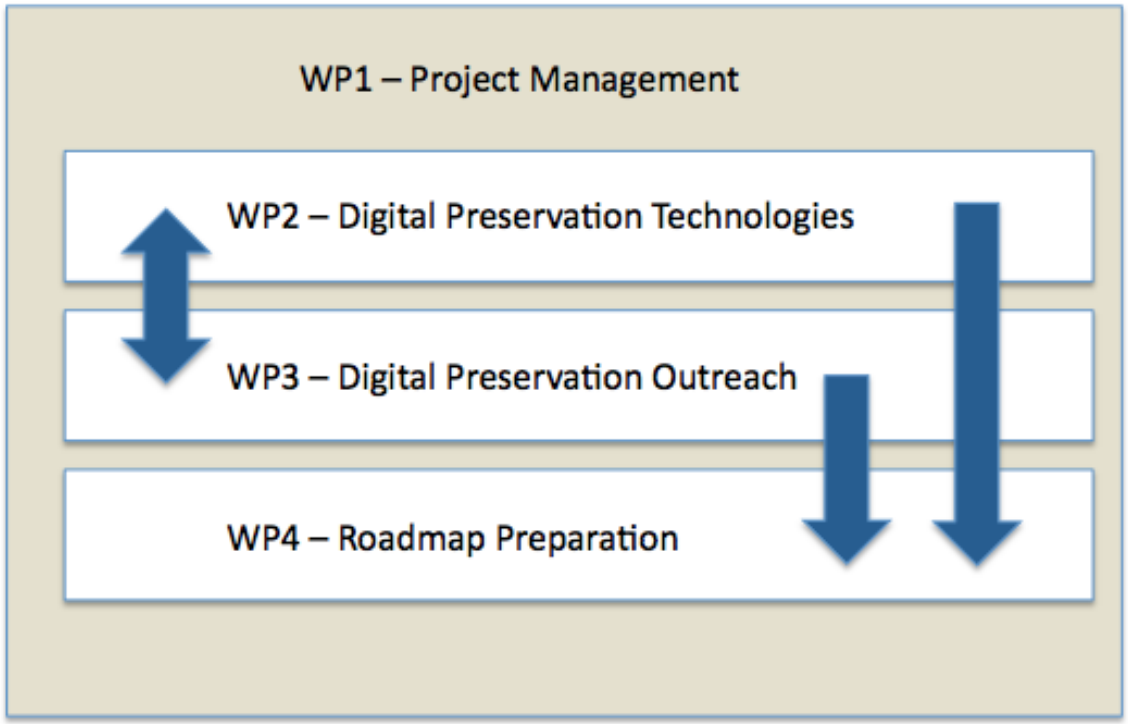
The motivation for and benefits from digital preservation would be widely presented to as many disciplines as possible, both to increase awareness of this matter and to seek further input and requirements for future activities. Existing networks and contacts would be used for this purpose and further ones established as necessary;

PODDS strategy -3

Building on this work, areas of key concern and urgent future work would be identified and prioritised with the explicit approval of the disciplines concerned;

A roadmap document would be drafted, widely disseminated and revised, seeking the approval of as many representative bodies and communities, such as the EIROforum members.

Work Packages

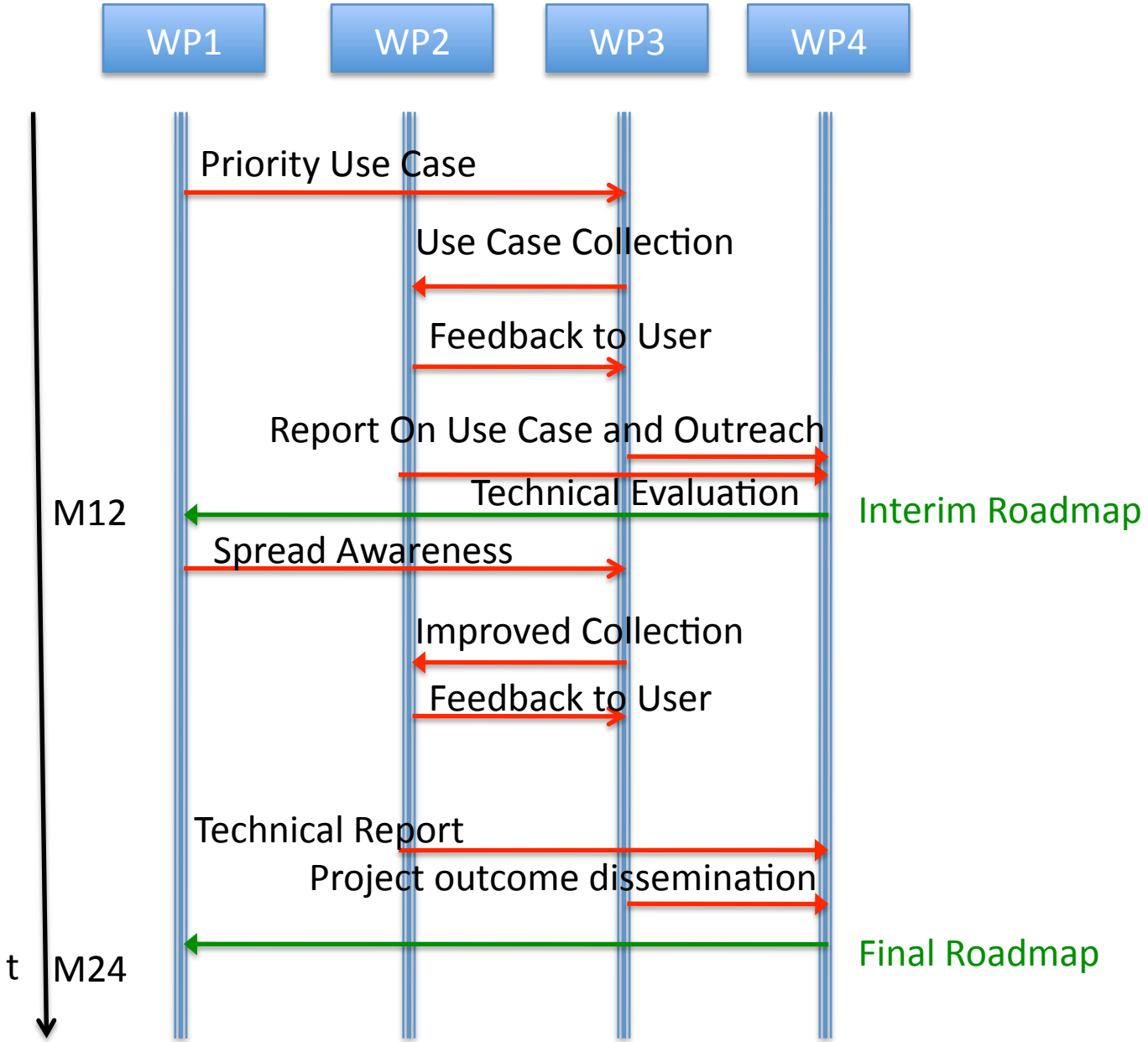


INFN

MPG

CERN

Work Flow


WP1
WP2
WP3
WP4
Priority Use Case
Use Case Collection
Feedback to User
Report On Use Case and Outreach
Technical Evaluation
Interim Roadmap
M12
Spread Awareness
Improved Collection
Feedback to User
Technical Report
Project outcome dissemination
Final Roadmap
t
M24

Use case example

C. Vuerli (INAF/IGI)

Astronomy & Astrophysics: BaSTI use-case

Basic Entities

- Stellar model simulations and their associated parameters
- Evolutionary FRANEC code (it feeds tracks and isochrones)
- Simulation output files and their associated metadata
- BaSTI Database
- Computing resources (DCIs)
- Web based portals

Goals

- Inspect BaSTI looking for tracks and isochrones and more generally for simulation output files
- Given a simulation output file, inspect BaSTI to retrieve information on the synthetic model simulation that produced the simulation output file
- Perform new simulations of existing synthetic models with a given configuration of the parameters space, and store the resulting simulation output file in BaSTI.
- Propose the ingestion of new Synthetic Models in BaSTI
- Perform post-processing analysis on the results of the simulation output files

OAIS

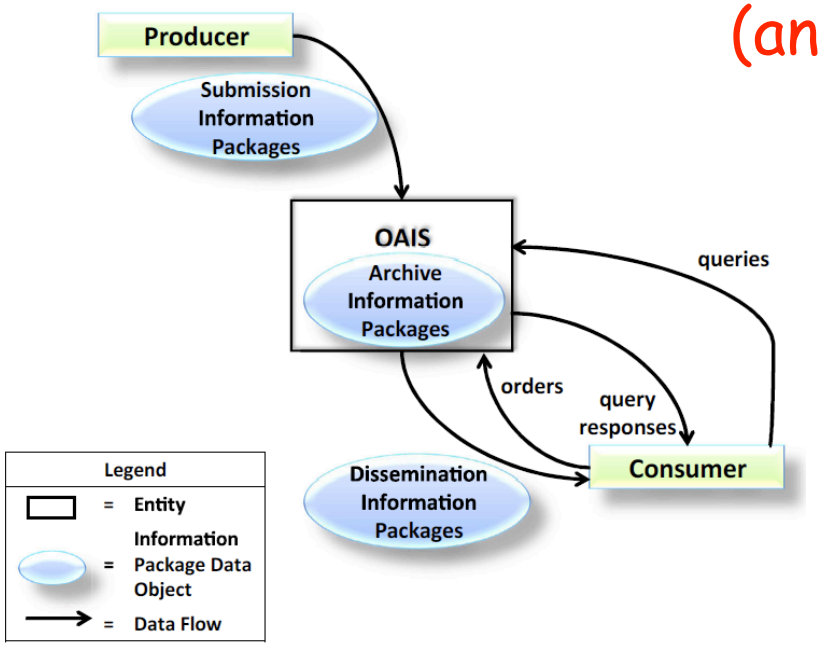
Consultative Committee for Space Systems

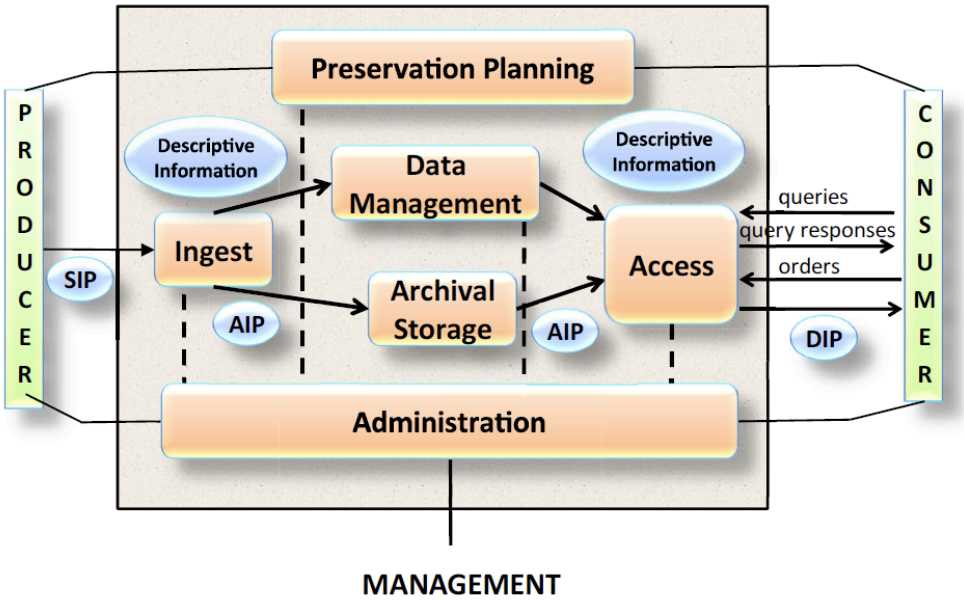
Recommendation for Space Data System Standards

REFERENCE MODEL FOR AN

Open Archival Information System

(an ISO)





Common DP issues

Long term organisation

Authorship

Supervision of the data preservation process

Access to and use of preserved data (VM, Emulators, frozen systems)

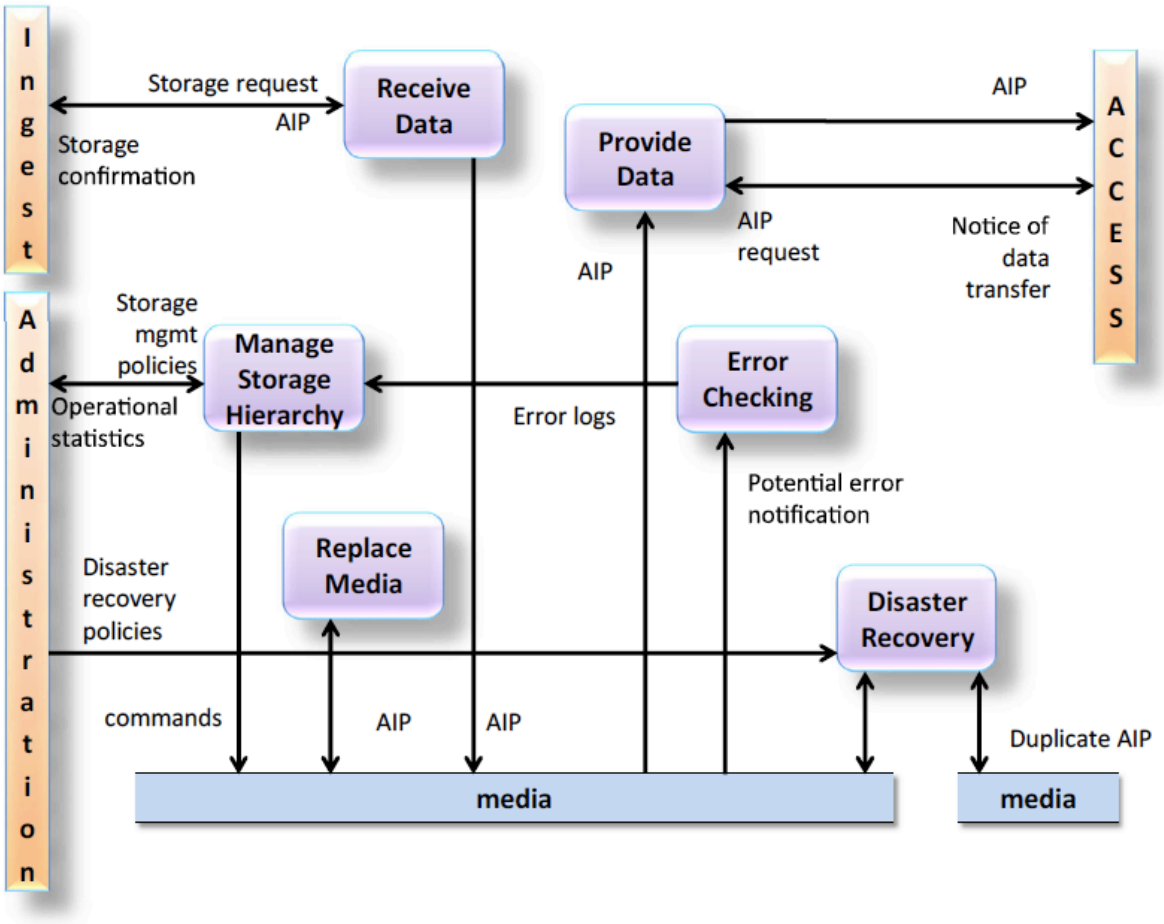
Preservation formats, standards

Storage, Data size and data integrity

Data Discovery

Validation

OAIS Archival Storage



http://fits.gsfc.nasa.gov/

Home | News | Docs | WCS | Samples | Libraries | Viewers | Utilities | Keywords | Conventions | Resources

The FITS Support Office

at NASA/GSFC

What is FITS?

- The standard data format used in astronomy
- Stands for 'Flexible Image Transport System'
- Endorsed by NASA and the International Astronomical Union
- Much more than just another image format (such as JPEG or GIF)
- Used for the transport, analysis, and archival storage of scientific data sets
  - Multi-dimensional arrays: 1D spectra, 2D images, 3D+ data cubes
  - Tables containing rows and columns of information
  - Header keywords provide descriptive information about the data
- See also the Wikipedia entry

See M. Nanni talk yesterday

INFN Impact

- Increasing INFN/HEP interest, coming from experiments
- Territorial distribution would help the leverage of regional/national projects
- Preparation to HORIZON 2020 through PODDS, if approved, but also independently

IGI support for infrastructure and scientific network