Data Management Plan

E. Fioretto

INFN – Laboratori Nazionali di Legnaro

Outline

- Introduction:
  - What is “Research data”?
  - Why should we preserve and share data?
- Life cycle of research data
- Research Infrastructures and Open Data
- Ongoing initiatives on Open Science and Data
- Data Management Plan for INFN National Laboratories
- Summary

GenOA week 2024
International Open Access week

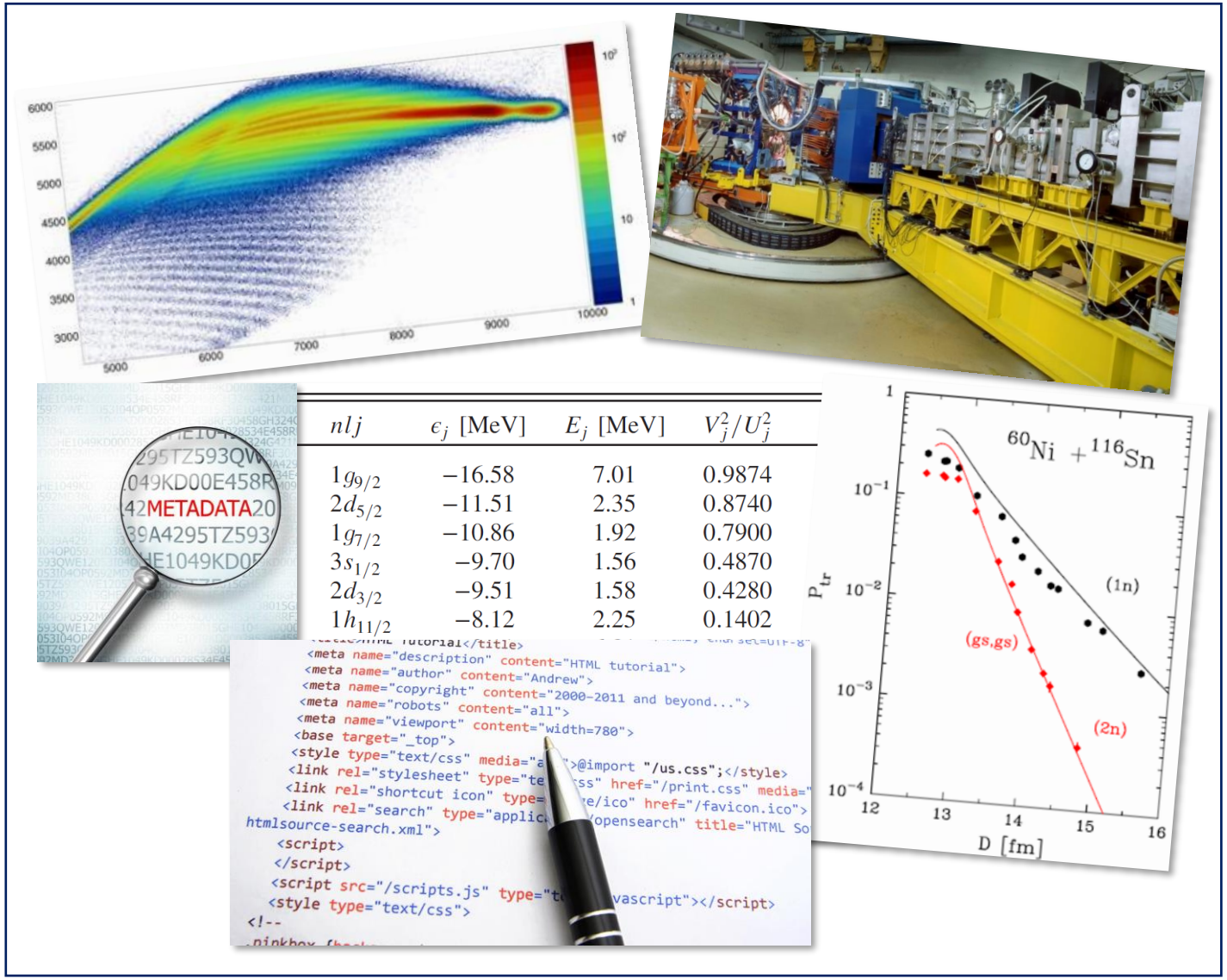
05/11

Comunità INFN e CoPER
- La scienza aperta per gli EPR: fare community
- La strada INFN verso l'Open Science

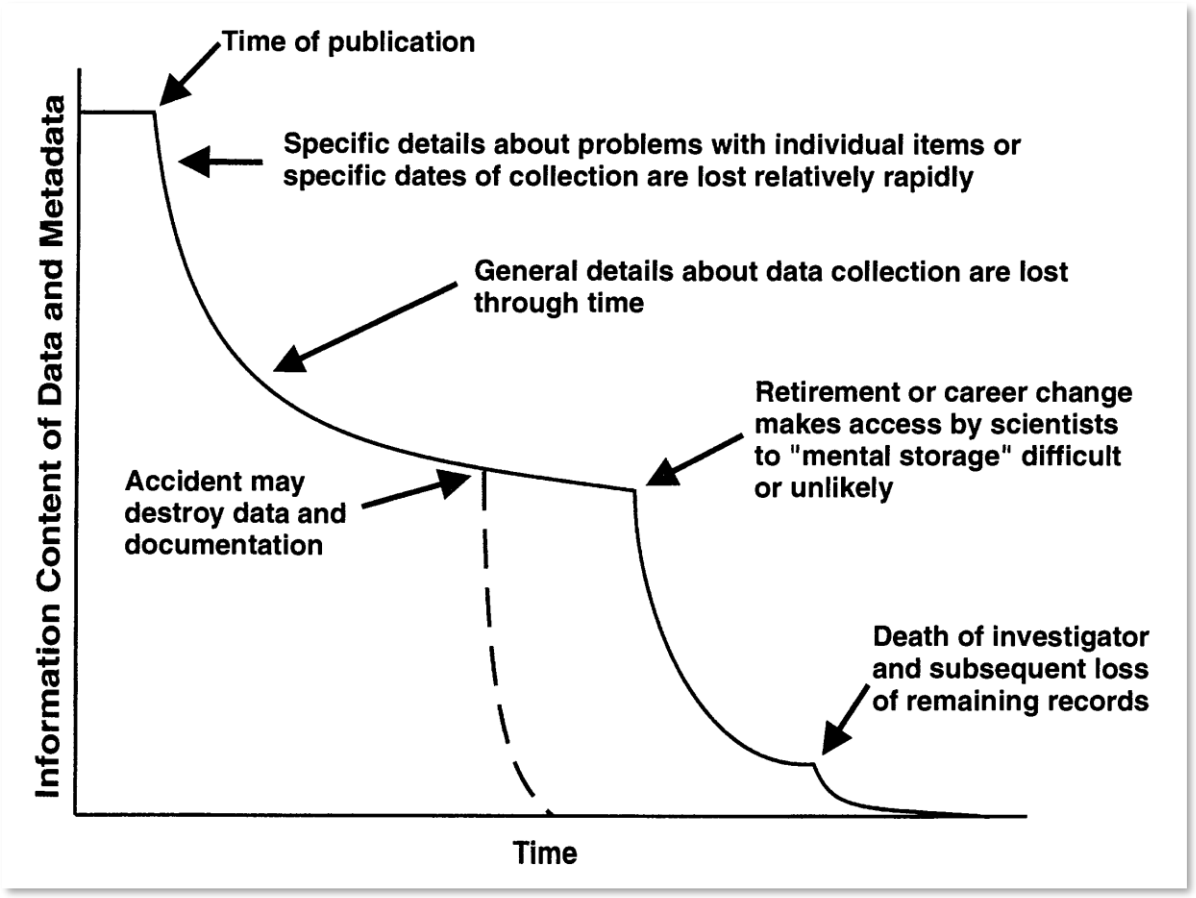
What is "Research data"?

- ✓ **Experimental datasets**
  - Raw data
  - Preprocessed data
  - Metadata
- ✓ **Simulations, results of calculations**
- ✓ **Software**
  - Source codes
  - Workflows
- ✓ **Databases**
- ✓ **Documentation**
  - All scientific documents related to the dataset (proposal, thesis, publications, manuals, notes, etc.)

ALL DIGITAL OBJECTS COLLECTED AND PRODUCED DURING A RESEARCH PROJECT



WHY SHOULD WE PRESERVE AND CURATE RESEARCH DATA?



W. K. Michener et al., Eco. App. 7 (1997) 330-342

- ✓ **The existing digital ecosystem surrounding the publication of research data prevents us from extracting maximum benefit from research investments**
- ✓ **Publications typically preserve some of the metadata**, but often only a subjectively selected portion of the metadata needed to relate the data to a specific hypothesis
- ✓ The **human memory is short** and many details on the data taking can be lost through time
- ✓ Moreover, the scenario is complicated by:
  - the extreme complexity of research data whose reanalysis requires a **complete set of metadata**
  - **accidents in storage media** through catastrophic events which may destroy data and metadata
  - **personnel turnover**
  - ……

WHY SHOULD WE PRESERVE AND SHARE RESEARCH DATA?

It is a duty towards funding bodies

- UE Grant Agreements
- PNRR Grants
- USA Grant

It is a duty towards funding foundations

- National Foundations (Cariparo)
- International Foundations

- Science funders, publishers and government agencies are increasingly requiring data management and stewardship plans for research data generated in publicly funded experiments
- In addition to the proper collection and storage, data management should also include the notions of "long-term preservation and curation" of data, with the aim of making them reusable for subsequent investigations, alone or in combination with newly collected data.

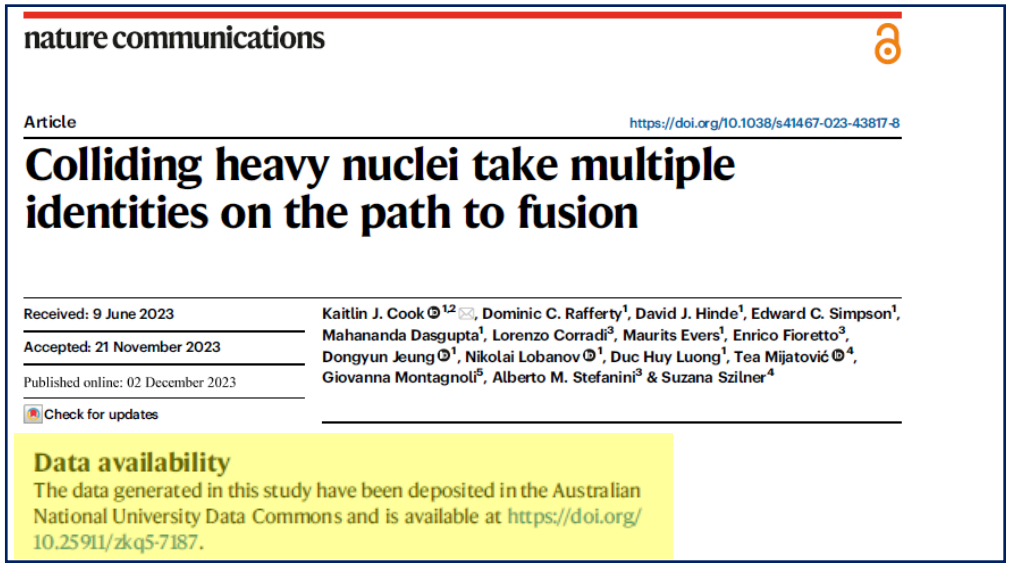
SPRINGER NATURE

Research data policy

At Springer Nature we advance discovery by publishing trusted research, supporting the development of new ideas and championing open science. We also aim to facilitate compliance with research funder and institution requirements to share data.

To help accomplish this we have established a standard research data policy for our journals, based on transparency around supporting data. This policy applies to all datasets that are necessary to interpret and replicate the conclusions reported in a research article.

1. All original articles must include a data availability statement



nature communications

Article

https://doi.org/10.1038/s41467-023-43817-8

Colliding heavy nuclei take multiple identities on the path to fusion

Received: 9 June 2023

Accepted: 21 November 2023

Published online: 02 December 2023

Check for updates

Kaitlin J. Cook[1,2], Dominic C. Rafferty[1], David J. Hinde[1], Edward C. Simpson[1], Mahananda Dasgupta[1], Lorenzo Corradi[3], Maurits Evers[1], Enrico Fioretto[3], Dongyun Jeung[1], Nikolai Lobanov[1], Duc Huy Luong[1], Tea Mijatović[4], Giovanna Montagnoli[5], Alberto M. Stefanini[3] & Suzana Szilner[4]

Data availability

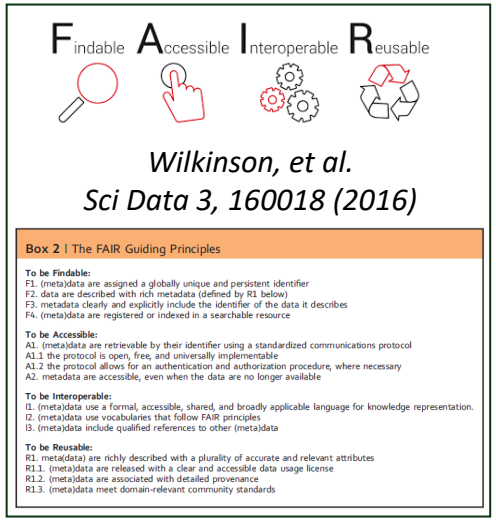
The data generated in this study have been deposited in the Australian National University Data Commons and is available at https://doi.org/10.25911/zkq5-7187.

ELSEVIER

The following principles underpin Elsevier's research data policy:

- Research data should be made available free of charge to all researchers wherever possible and with minimal reuse restrictions

Life cycle of Research Data



Re-using data

- ✓ Follow-up research
- ✓ New research
- ✓ Undertake research reviews

Taking data

- ✓ Design experiment (proposal, DOI, etc.)
- ✓ Collect data
- ✓ Capture and create metadata

Processing data

- ✓ Check and validate data
- ✓ Manage and storage all auxiliary data

Giving access to data

- ✓ Share data
- ✓ Access control
- ✓ Establish copyright

Analyzing data

- ✓ Interpret and derive data
- ✓ Produce research outputs (publications, reports, etc.)
- ✓ Prepare data for preservation

Preserving data

- ✓ Migrate data to suitable medium
- ✓ Back-up and store data
- ✓ Store metadata and documentation

Research Infrastructures and Open Data

How can Research Infrastructures (RI) advance Open Science by providing pertinent services for the Data Management?

- **All RI are Data Producers**
- **All RI are Science Driver**
- **All RI have a responsibility on the management of collected data**
  - FAIR principles at every stage of experiments planning
  - Long term preservation of datasets
  - Openness of the data (for scientific communities and citizens)
  - Re-usability

A DMP of the RI is a good starting point to formalize and make explicit the data management strategy

Ongoing initiatives on Open Science and Data

NuPECC Long Range Plan 2024

Open Science and Data Thematical Working Group

Members:

- Hector Alvarez-Pol, USC, Spain
- Stefano Bianco, INFN Frascati, Italy
- Vivian Dimitriou, IAEA, Austria
- Xavier Espinal, CERN, Switzerland
- Michel Jouvin, CNRS/IJCLab, France
- Adrien Matta, CNRS/LPC Caen, France
- Caterina Michelagnoli, ILL, France
- Andrew Mistry, GSI/FAIR, Germany
- Panu Rahkila, JYFL, Finland
- Manuela Rodriguez, Sevilla, Spain
- Olivier Stezowsky, CNRS/IP2I, France
- Enrico Vigezzi, INFN Milano, Italy

NuPECC LRP2024 Report
Presentation Meeting in Brussels
November 19, 2024

GOALS

- Provide efficient transnational access to the available resources at a major fraction of EUROpean Laboratories for Accelerator Based Sciences (EURO-LABS) at a network including the major European laboratories
- Enhance collaborative targeted improvements for the existing services that will lead to an increase of the scientific and technical opportunities at various Ris
- **Make the results** from the tests conducted **at the RIs of EURO-LABS during the period of the project freely available to the scientific community and manage the experimental data**, when relevant, **through a Data Management Plan (DMP) in line with the FAIR principles**
- Organize the training of the new generation of researchers and young technical staff to best exploit the RIs, through workshops and hands-on experience at specifically chosen RIs

To comply with the Grant Agreement: Experiments receiving support from TNA should follow DMP from RI. What if RI do not have a DMP?

Deliverable 5.7 M6
Release of the initial Data Management Plan of the Project

DMP for INFN National Laboratories

A Working Group was established within CSN3 to discuss and propose a draft DMP for INFN National Laboratories.

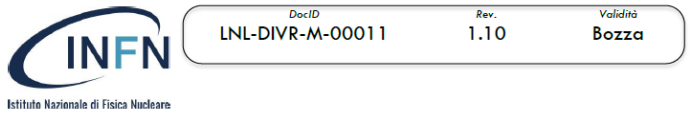
WG members
E. Fioretto – LNL (Coordinator)
F. Ferraro & F. Marchegiani – LNGS
S. Pisano – LNF
M. La Cognata – LNS
in collaboration with WG Open Science of INFN

Different kinds of experiments are performed at INFN National Laboratories:

- **Data taking at particle accelerators** (CSN1, CSN3 and CSN5)
- **Data taking at astroparticle physics facilities** (CSN2)
- **Characterization of detectors, electronics, and materials** (CSN1, CSN2, CSN3 and CSN5)
- **Interdisciplinary physics activities** (CSN5)

DRAFT DMP FOR LNL AND LNS

A Data Management Plan (DMP) is the formal document that describes the management of research data to be collected, processed and/or generated by an experiment.



LNL Data Management Plan

Written by	Checked by	Approved by
Enrico Fioretto	Enrico Fioretto Tommaso Marchi (EURO-LABS TNA coordinator) Javier Valiente-Dobon (Gr3 Coordinator) Valeria Conte (Gr5 Coordinator) Antonello Ortolan (Gr2 Coordinator) Michele Gulmini (Responsible for the Information Technology Service)	

Subject

Data Management Plan for research activities making use of LNL ion beams.

LNS Data Management Plan

Written by	Checked by	Approved by
M. La Cognata (CSN3 coordinator)	S. Gammino (Director) A. Tumino (Head of the research division) V. Greco (Deputy director) E. Giorgio (Responsible for the IT Service) G. Riccobene (CSN2 coordinator) D. Gambacurta (CSN4 coordinator) G. Torrisi (CSN5 coordinator)	

Subject

Data Management Plan for research activities making use of Laboratori Nazionali del Sud (LNS) facilities.

DEFINITION OF TERMS

- SPOKESPERSON: Person responsible for the experiment, identified on the scientific proposal submitted to the Program Advisory Committee, and for the data management collected during the experiment.
- **DATA STEWARD: Person responsible for the management of research data throughout their life cycle, from the collection phase to the storage and sharing ones**.
- EXPERIMENTAL ACCOUNT: Directory created by the Information Technology Service containing the data record for an experiment and therefore all the data linked to the experiment.
- **DIGITAL OBJECT IDENTIFIER** (DOI): **Unique, long-term identifier allowing the identification of a dataset**. This identifier will be created by Open Science Working Group as routinely done for the sharing of products on the Open Access Repository (OAR).
- **EMBARGO PERIOD: Period during which the data are available only to the experimental team. Beyond that period the data must be open to the widest audience**.
- EXPERIMENT CONTACT PERSON: Local staff member who facilitates the running of the experiment.
- RAW DATA: All kinds of data collected by experiments carried out by using ion beams delivered by LNL accelerators.
- METADATA: All information necessary to manage and perform the analysis of the raw data, including (but not limited to) the context of the experiment, the experimental team, the experimental conditions, the data format, the logbook, software package, etc.
- BEAMTIME COORDINATOR: Person in charge of the coordination of the accepted experiments and the preparation of the beamtime schedule maximizing the number of experiments to be performed in the available accelerator beamtime.

OWNERSHIP OF DATA

INFN is the owner and the custodian of the raw data (and associated metadata) produced by using the instrumentation installed in its National Laboratories.

Often, **large collaborations have already a DMP or are ruled by international agreements such as MoUs**. **In such cases specific agreements between the INFN laboratories and the Management Boards of the collaborations have to be established**.

All raw data (and associated metadata) collected in experiments approved by the Program Advisory Committee (excluding commercial use of the research infrastructures) **will be open access after an initial embargo period during which access is restricted to the experimental team, represented by the spokesperson**.

All raw data (and associated metadata) obtained as a result of proprietary research will be owned exclusively by the client who purchased the beamtime and is not covered by this DMP. Commercial users must agree with the facility management on how they wish their raw data and metadata to be managed before the start of any experiment.

Curation of Raw Data and associated Metadata

Raw data and metadata will have read-only access for the duration of their life cycle.
Raw data formats must be well documented in the metadata.
All raw data and metadata will be organized in a well-defined structure which will be made available by INFN Laboratories. **Only raw data with associated metadata will be archived**.

The spokesperson has to inform data steward about the requirements in terms. e.g., of disk space upon the scheduling of the experiment. **The spokesperson has the responsibility to provide the data and the metadata (in electronic or pdf format) to the data steward, in compliance with the FAIR principles**. The spokesperson has to ensure that experiments' metadata are complete, as this will enhance the possibilities for everybody to search for, retrieve and interpret the data in the long term.

Each experiment and dataset will have a unique permanent Digital Object Identifier (DOI). This DOI will be assigned by the Open Science Working Group as routinely done for the sharing of INFN products on the Open Access Repository (OAR).
Anybody publishing results based on open access data must quote the same identifier.

Access to Raw Data and Metadata

Raw data and metadata will be stored on a short-term basis (one year maximum) in dedicated local servers and preserved for immediate access and data analysis.

After that, raw data and metadata will be transferred and stored to INFN-CNAF and preserved for at least fifteen years, unless differently agreed (for instance, in the case of already existing DMP).

High level metadata such as Title, Authors, Abstract, will be made public as soon as possible using a dedicated webpage (https://opendata.xxxx.infn/).
This information will be available through the persistent identifier.
A data steward will be in charge of the curation of the data as specified in the present document.

Publications' Information

Publications related to data collected in experiments performed at INFN National Laboratories must cite the persistent identifier of the data.

Data Management Strategy

Scientific use of the research infrastructures

Submission of the scientific proposal

DMP acceptance

PAC approval

DOI assignment

Acceptance of this policy is a necessary condition for the awarding of the beam time at accelerators

Data taking

Resident detectors

Itinerant detectors

A copy of raw data and metadata must be provided by the spokesperson

Short –term storage of data and metadata on dedicated local servers

For data analysis

Immediate access

Preserved 1 year maximum

Data organization

- «Identifier» DATA
  - Raw data
- METADATA
  - eLog
  - Exp. Minutes
  - Logbook
- TEMP
- MISC
  - Processed data
  - Analysis
  - Software

Long-term storage @ INFN-CNAF dedicated servers

Embargo period at least 5 years

Access upon request

Preserved at least 15 years

Commercial use of the research infrastructures

Submission of the request

Director approval

Data taking

Commercial use of infrastructures
Outside/local DAQ and storage

SUMMARY

- **A DMP for INFN National Laboratories is necessary and useful**
  - Allow to collect and store data and metadata in a more structured way
  - Avoid or minimise risk of data loss
  - Enable share/re-use of data and guarantees research reproducibility
  - Increase verifiability of research
  - Increase longevity of data by helping to make them available even after project ends

- **The release of an initial DMP for LNL and LNS is required within the EU Grant Agreement EURO-LABS**

- **Draft documents have already been written**

- **Additional resources are needed for an Open Data Policy**
  - Personnel (Data Steward at INFN Laboratories)
  - Hardware (Storage servers at CNAF or other Computing Centres)

- **All CSN of INFN should be involved in the discussion and the drafting the DMPs**