Data management and Open Science
EURO-LABS WP 5.2

Outline

- Foreword : Digital Objects and (Research) Data Management
- EURO-LABS Data Management Plan
- WorkPackage 5.2 Open Science and Data Tasks

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101057511.

Foreword : Digital Objects

Data are any digital objects :

- Experimental Data-sets
  (raw, auxiliary-data, refined, ...)
- Simulations, Results of calculations
- Databases
- Software (sources code, Workflows, ...)
- Reports, Publications, Slide-Shows, Websites, Pictures, ...
- Data Management Plans !
- ...

Foreword : Digital Objects

Data are any digital objects :

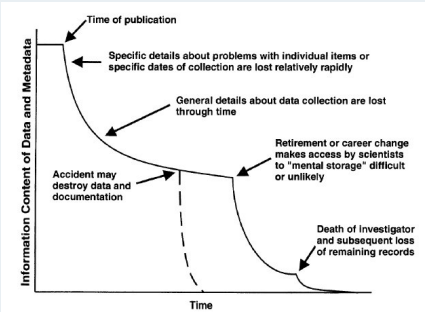
- Experimental Data-sets
  (raw, auxiliary-data, refined, ...)
- Simulations, Results of calculations
- Databases
- Software (sources code, Workflows, ...)
- Reports, Publications, Slide-Shows, Websites, Pictures, ...
- Data Management Plans !
- ...

Data have a Life Cycle !

Picture from Research Data Lifecycle by LMA Research Data Management Working Group

Foreword : Why should we care about Data Management ?

Data and Metadata Entropy



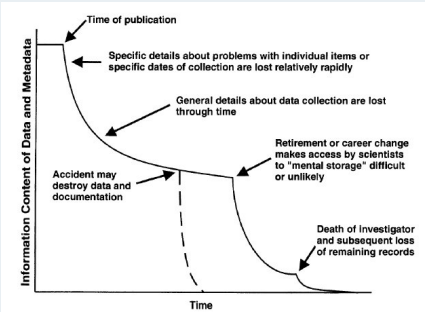
W. K. Michener et al., Eco. App. 7 (1997) 330-342

No data set is perfect and self-explanatory

- Too often relying on human/mental storage
- Crucial to accurately interpret results and their origin (from processing, analysis, and modeling)
- Accessibility and Reproducibility of research results
- Enhance visibility of research within and outside research domain

Foreword : Why should we care about Data Management ?

Data and Metadata Entropy



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

No data set is perfect and self-explanatory

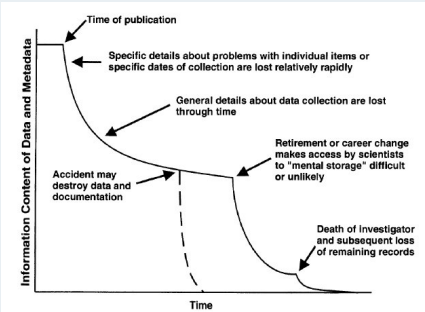
- Too often relying on human/mental storage
- Crucial to accurately interpret results and their origin (from processing, analysis, and modeling)
- Accessibility and Reproducibility of research results
- Enhance visibility of research within and outside research domain

Long term Preservation and Management

- Defining data policies (access, sharing, preservation period, ...)
- Re-Use opportunities, Facilitate Cross Domain research
- How to chose if a data-set should be kept (Unlimited storage erea is behind us !)

Foreword : Why should we care about Data Management ?

Data and Metadata Entropy



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

No data set is perfect and self-explanatory

- Too often relying on human/mental storage
- Crucial to accurately interpret results and their origin (from processing, analysis, and modeling)
- Accessibility and Reproducibility of research results
- Enhance visibility of research within and outside research domain

Long term Preservation and Management

- Defining data policies (access, sharing, preservation period, ...)
- Re-Use opportunities, Facilitate Cross Domain research
- How to chose if a data-set should be kept (Unlimited storage erea is behind us !)

How to address these challenges?

It is now mandatory for proposals/projects to include Data Management Plan.

Foreword : FAIR Data Practices

Findable :

- F1 Data are assigned a globally unique and persistent identifier
- F2 Data are described with rich meta data
- F3 Meta data clearly and explicitly included the identifier of the data they described
- F4 (Meta)data are registered or indexed in a searchable resource

Accessible :

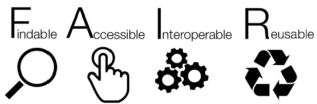
- A1 (Meta)data are retrievable by their identifier using a standardized communication protocol
- A1.1 The protocol is open, free and universally implementable
- A1.2 The protocol, where necessary, allows for an authentification & authorisation procedure
- A2 Metadata are accessible, even when the data are no longer available

Inter-operable :

- I1 (Meta)data use a normal, accessible, shared and broadly applicable language for knowledge representation
- I2 (Meta)data use vocabularies that follow FAIR principles
- I3 Meta-data qualified references to other (meta)data

Reusable :

- R1 (Meta)data are richly described with a plurality of accurate and relevante attributes
- R1.1 (Meta)data are released with a clear and accessible usage licence
- R1.2 (Meta)data are associated with detailed provenance
- R1.3 (Meta)data meet domain-relevant community standards



Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)

Foreword : FAIR Data Practices

Findable :

- F1 Data are assigned a globally unique and persistent identifier
- F2 Data are described with rich meta data
- F3 Meta data clearly and explicitly included the identifier of the data they described
- F4 (Meta)data are registered or indexed in a searchable resource

Accessible :

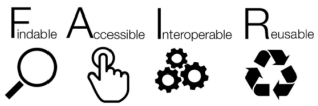
- A1 (Meta)data are retrievable by their identifier using a standardized communication protocol
- A1.1 The protocol is open, free and universally implementable
- A1.2 The protocol, where necessary, allows for an authentification & authorisation procedure
- A2 Metadata are accessible, even when the data are no longer available

Inter-operable :

- I1 (Meta)data use a normal, accessible, shared and broadly applicable language for knowledge representation
- I2 (Meta)data use vocabularies that follow FAIR principles
- I3 Meta-data qualified references to other (meta)data

Reusable :

- R1 (Meta)data are richly described with a plurality of accurate and relevante attributes
- R1.1 (Meta)data are released with a clear and accessible usage licence
- R1.2 (Meta)data are associated with detailed provenance
- R1.3 (Meta)data meet domain-relevant community standards



Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)

How FAIR are we/you ?

Required critical exercice of our/your level of data FAIRness

Foreword : FAIR Data Practices

Findable :

- F1 Data are assigned **a globally unique and persistent identifier**
- F2 Data are **described with rich meta data**
- F3 Meta data clearly and explicitly included the identifier of the data they described
- F4 (Meta)data are registered or indexed in a searchable resource

Accessible :

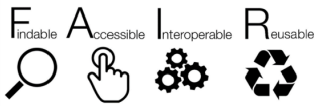
- A1 (Meta)data are **retrievable by their identifier** using a standardized communication protocol
- A1.1 The protocol is open, free and universally implementable
- A1.2 The protocol, where necessary, **allows for an authentification & authorisation procedure**
- A2 **Metadata are accessible, even when the data are no longer available**

Inter-operable :

- I1 (Meta)data use **a normal, accessible, shared and broadly applicable language for knowledge representation**
- I2 (Meta)data use vocabularies that follow FAIR principles
- I3 Meta-data qualified references to other (meta)data

Reusable :

- R1 (Meta)data are **richly described with a plurality of accurate and relevante attributes**
- R1.1 (Meta)data are released with a clear and accessible usage licence
- R1.2 (Meta)data are associated with detailed provenance
- R1.3 (Meta)data **meet domain-relevant community standards**



Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)

How FAIR are we/you ?

Required critical exercice of our/your level of data FAIRness

Foreword : Data and Open Science

Ambitions of Open Science

- Change the way citizens perceive research and public investment in research
- Enable opportunities offered by digital revolution to allow everybody to contribute to the scientific process.
- Accelerate discoveries and increase the scientific value by sharing and transferring knowledge within and between scientific communities
- Contribute to training of next generation of researchers

Foreword : Data and Open Science

Ambitions of Open Science

- Change the way citizens perceive research and public investment in research
- Enable opportunities offered by digital revolution to allow everybody to contribute to the scientific process.
- Accelerate discoveries and increase the scientific value by sharing and transferring knowledge within and between scientific communities
- Contribute to training of next generation of researchers

Open Data means "FAIR" Data

- Endorse the "FAIR" principles (Findable, Accessible, Interoperable, Reusable)
- "Open data" does not mean "free data"
  Access Management, Data policies, ...
  $\rightarrow$ properly managed digital objects !

European Open Science Cloud (EOSC) and scientific clusters

European Open Science Cloud

- EOSC is the EU action in response to EU member states shared policy about the uptake of Open Science
- EOSC is a cloud for research data in Europe that allows universal access to data
- EOSC is federating existing resources across Europe (national data center, research infrastructures, ...) to allow researchers and citizens to access and re-use data produced by other scientists.

European Open Science Cloud (EOSC) and scientific clusters

European Open Science Cloud

- EOSC is the EU action in response to EU member states shared policy about the uptake of Open Science
- EOSC is a cloud for research data in Europe that allows universal access to data
- EOSC is federating existing resources across Europe (national data center, research infrastructures, ...) to allow researchers and citizens to access and re-use data produced by other scientists.

Scientific clusters

Scientific clusters related to Nuclear Physics and HEP: developing and providing services for scientific data and connecting these to the European Open Science Cloud (EOSC).

- ESCAPE → ESFRI facilities in astronomy- and accelerator-based particle physics + nuclear physics (GSI/FAIR)
- PANOSC → Photon and Neutron Open Science Cloud

Where are we standing on Research Data Management ?

Context and Challenges [in Nuclear Physics]

- Increasing volume of data produced → Volumetry
- Increasing complexity of the data produced → Metadata / Workflows
- Increasing size of collaboration → Colloborative analyses
- Heterogeneous situations of Data Producers :
  - Research Infrastructures and their local instruments
  - Collaboration for Travelling Detectors
- Heterogeneous [Nuclear Physics] Communit(ies)
  National/International large size collaborations ↔ smaller size research groups

ENSAR2 SATNuRSE JRA - D13-3 "Inventory and protocol for data management" (28.02.21)

Given that most respondents see advantages, we conclude that it is time to start the design and development of a European data-storage system using FAIR principles. This system must be capable to store the raw data from the detectors and in addition metadata to allow another researcher to analyze the data.

Where are we standing on Research Data Management and Open Science ?

Ongoing initiatives

Data Management Plans / Data Policies

- National/Local initiatives on Data Management Plans and Policies (requirement)
  GSI/FAIR, GANIL, JYFL, CERN Data Policy and Open Data WG ISOLDE/CERN, AGATA ...
- **Not yet everywhere** and heterogeneous (possible conflicting policies between Travelling detectors and RI)

Catalogues and Data Storage Practices:

- National/Local initiatives from RI on catalogues (JYFL, GANIL, ...)
- Storage : from local disk to data centers !
- Limitations on data-set access (techical or polycy-wise)
- MetaData is definitevely the poor's man in data management (Physics, Detectors, Accelerator)

Softwares :

- Signifiant move towards open softwares (FAIRroot, GammaWare, nptool, Kaliveda, ...)

Structurating Data Management towards Open Science

Research Infrastructures have a key role in the structuration and convergence Open Science practices,
but this will only be possible with the inputs of the community and a coordination between RIs to prevent policy conflicts.

How Research Infrastructures can advance Open Science by providing pertinent services for Data and Software Management ?

- All RI are Data producer
- All RI are Science Drivers

How Research Infrastructures can advance Open Science by providing pertinent services for Data and Software Management ?

- All RI are Data producer
- All RI are Science Drivers
- All RI Have a "responsability" on the management of collected data :
  - FAIR principles at every stages of experiments planning
  - Long term preservation of datasets
  - Openness of the data (for scientific communities and citizens),
  - Re-usablity

How Research Infrastructures can advance Open Science by providing pertinent services for Data and Software Management ?

- All RI are Data producer
- All RI are Science Drivers
- All RI Have a "responsability" on the management of collected data :
  - FAIR principles at every stages of experiments planning
  - Long term preservation of datasets
  - Openness of the data (for scientific communities and citizens),
  - Re-usablity

DMP are a good starting point to formalize and make explicit challenges and answers

How to improve data management practices towards Open Science?

Data Management and Preservation

- FAIR Data Principles (DMPs, ...)
- Data Catalogs
- Enhanced and automated Meta Data collection (Physics, Detection, Accelerator)
- Standard unpacking software
- Authentification Authorisation Infrastructure AAI
- Data Portal Platforms

Softwares and Analysis Management

- Analysis Software Catalogs
- Data Workflow Catalogs

Analysis Platforms

- For scientists (collaborative analysis, virtual organisations around scientific topic ...)
- For training next generation of scientists

How to improve data management practices towards Open Science?

Data Management and Preservation

- FAIR Data Principles (DMPs, ...)
- Data Catalogs
- Enhanced and automated Meta Data collection (Physics, Detection, Accelerator)
- Standard unpacking software
- Authentification Authorisation Infrastructure AAI
- Data Portal Platforms

Softwares and Analysis Management

- Analysis Software Catalogs
- Data Workflow Catalogs

Analysis Platforms

- For scientists (collaborative analysis, virtual organisations around scientific topic ...)
- For training next generation of scientists

A very long way to go !

→ Generalize the adoption of open data policies, standard metadata and FAIR data stewardship

→ Establish the basis and framework for a workable Open Science initiative that will accelerate discoveries, maximize scientific value by sharing data and knowledge within scientific communities and would allow nuclear physics to broaden its impact in both science and society.

How to improve data management practices towards Open Science?

Data Management and Preservation

- **FAIR Data Principles (DMPs, ...)**
- **Data Catalogs**
- Enhanced and automated Meta Data collection (Physics, Detection, Accelerator)
- Standard unpacking software
- **Authentification Authorisation Infrastructure AAI**
- **Data Portal Platforms**

Softwares and Analysis Management

- Analysis Software Catalogs
- Data Workflow Catalogs

Analysis Platforms

- For scientists (collaborative analysis, virtual organisations around scientific topic ...)
- For training next generation of scientists

A very long way to go !

→ Generalize the adoption of open data policies, standard metadata and FAIR data stewardship

→ Establish the basis and framework for a workable Open Science initiative that will accelerate discoveries, maximize scientific value by sharing data and knowledge within scientific communities and would allow nuclear physics to broaden its impact in both science and society.

→ At the core of EURO-LABS WP 5.2 actions/deliverables

EURO-LABS WP 5.2:

Participants

- CSIC, GANIL (Leading partners), INFN, CNRS, IJCLab, GSI
- Task leader: A. Lemasson (Ganil);
- Subtasks leader : C. Hornung (GSI), A. Matta (CNRS/LPC Caen), M. Jouvin (CNRS/IJCLab)

Goals

- Bringing the nuclear physics community into the EOSC (European Open Science Cloud) framework
- Developing services to enhance FAIR (Findable, Accessible, Interoperable and Reusable) data principles
- Integration of Nuclear Physics community to existing infrastructures/services of EOSC environment - benefiting from the experience from ESCAPE/HEP physics community

Work Package Structure

- Task 1 : EURO-LABS Data Management Plan + Open Science Desk for RI
- Task 2 : Data Catalog : Open NP
- Task 3-1 : Authorization and Authentification Plateform
- Task 3-2 : Prototype of Data Access Platform

Task 1 : EURO-LABS Data Management Plan

Task Leaders : CSIC (M. J. G. Borge) / GANIL (A. Lemasson)

Data Management Plan of Data Produced with EURO-LABS Project

- Each WP/task Leader have to identify the data that will be produced within EURO-LABS (Reports, Slides, Softwares, Video, ...)

Task 1 : EURO-LABS Data Management Plan

Task Leaders : CSIC (M. J. G. Borge) / GANIL (A. Lemasson)

Data Management Plan of Data Produced with EURO-LABS Project

- Each WP/task Leader have to identify the data that will be produced within EURO-LABS (Reports, Slides, Softwares, Video, ...)
- Zenodo plateform could be a good option for EURO-LABS digital objects (https://zenodo.org/) as used for several EU funded projects (AIDAinnova, ESCAPE, ...)

Task 1 : EURO-LABS Data Management Plan

Task Leaders : CSIC (M. J. G. Borge) / GANIL (A. Lemasson)

Data Management Plan of Data Produced with EURO-LABS Project

- Each WP/task Leader have to identify the data that will be produced within EURO-LABS (Reports, Slides, Softwares, Video, ...)
- Zenodo plateform could be a good option for EURO-LABS digital objects (https://zenodo.org/) as used for several EU funded projects (AIDAinnova, ESCAPE, ...)
- **To comply with Grant Agreement :**
  Experiments receiving support from TNA should follow DMP from RI
  What if RI do not have a DMP ?

Task 1 : EURO-LABS Data Management Plan

Task Leaders : CSIC (M. J. G. Borge) / GANIL (A. Lemasson)

Data Management Plan of Data Produced with EURO-LABS Project

- Each WP/task Leader have to identify the data that will be produced within EURO-LABS (Reports, Slides, Softwares, Video, ...)
- Zenodo plateform could be a good option for EURO-LABS digital objects (https://zenodo.org/) as used for several EU funded projects (AIDAinnova, ESCAPE, ...)
- **To comply with 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 **[!] Input Required**

Task 1 : Open Science Desk

Promoting and open science initiatives in RI

- Network of contact persons in each RI.
- Support for setting up Data Management Plan and Practices
- Regular discussion on development directions of Task 5.2.2 and 5.2.3
- Collect needs from RI users and members
- Provide feedback when plate-forms are released

Task 1 : Open Science Desk

Promoting and open science initiatives in RI

- Network of contact persons in each RI.
- Support for setting up Data Management Plan and Practices
- Regular discussion on development directions of Task 5.2.2 and 5.2.3
- Collect needs from RI users and members
- Provide feedback when plate-forms are released

Not a deliverable, but essential to tighten links between RI and Users

If you are interested in joining this activity, please contact us.

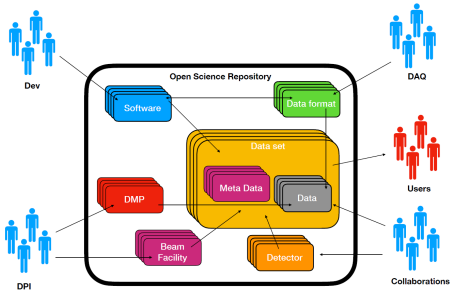
Task 2 : Open NP

Catalogue of data-sets, related information and tools

Task Leaders : LPC Caen (A. Matta) / GANIL (A. Lemasson)

EURO-LABS time scale goals :

- Catalogue of identified existing and future raw-data sets
- Catalogue of identified existing apparatus: ion sources, accelerator, separator and detector
- Associated aux-data (i.e. log book, ...)
- Associated software to exploit raw-data and aux-data
- Associated software to exploit and produce analyzed and simulated data



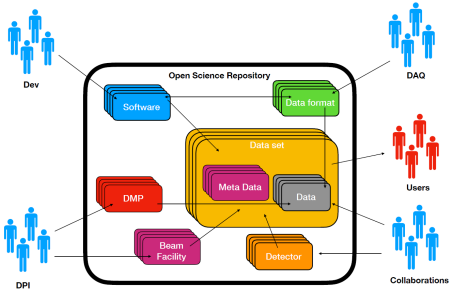
Task 2 : Open NP

Catalogue of data-sets, related information and tools

Task Leaders : LPC Caen (A. Matta) / GANIL (A. Lemasson)

EURO-LABS time scale goals :

- Catalogue of identified existing and future raw-data sets
- Catalogue of identified existing apparatus: ion sources, accelerator, separator and detector
- Associated aux-data (i.e. log book, ...)
- Associated software to exploit raw-data and aux-data
- Associated software to exploit and produce analyzed and simulated data



Deliverable 5.3 - M36

Release of the first functional version of the Open NP

Task 3-1 : Authentication and Identification Infrastructure

Task Leader : CNRS/IJCLab (M. Jouvin)

Global Authentication And Authorization Plateform

- support of heterogeneous authentication mechanisms
- abstraction of collaboration/virtual organization (membership management, access policies ...)
- delegation of privileges across the chain of services implementing the data lake vision
- integration in existing data access and computing software

Task 3-1 : Authentication and Identification Infrastructure

Task Leader : CNRS/IJCLab (M. Jouvin)

Global Authentication And Authorization Plateform

- support of heterogeneous authentication mechanisms
- abstraction of collaboration/virtual organization (membership management, access policies ...)
- delegation of privileges across the chain of services implementing the data lake vision
- integration in existing data access and computing software

A powerful tool for RI and the community beyond data management

- Identity management of theirs users and collaborations
- Rights management to access data
- Rights management to access services they propose
  (PAC submission, Remote Access, Theory Plateforms, ...)
- Identification Portal for collaborative tools of the community (Open-NP, Analysis Plate-forms, Virtual observatories., ...)

Task 3-2 : Prototype of data access portal

Devellop data access portal for users
Task Leader : GSI (C. Hornung)

Prototype

- Provide access to the data-set stored by RI to authenticated and authorized users
- Access Interfaces to anticipate integration in Data-Lake vision
- Benefiting from expertise from GSI (synergies with ESCAPE, PUNCH4NFDI, ...)
- Toolkit to be deployed in different RI.

Task 3-2 : Prototype of data access portal

Devellop data access portal for users
Task Leader : GSI (C. Hornung)

Prototype

- Provide access to the data-set stored by RI to authenticated and authorized users
- Access Interfaces to anticipate integration in Data-Lake vision
- Benefiting from expertise from GSI (synergies with ESCAPE, PUNCH4NFDI, ...)
- Toolkit to be deployed in different RI.

Deliverable D5.3 M36

Release of the first functional version of data access tools

Open Science in NP : a long journey ahead of us

Opportunities of for "Openned Science" Nuclear Physics

- Improve drastically management and visibility of data-sets
- Develop new collaborations based on combined or reused data-sets.
- Enhance the scientific impact of the available and future data-sets

Require strong involvement of all stakeholders (Researchers, RI, IT departments ...)
starting from having Data Management Plans

One step with EURO-LABS WP 5.2, towards an ambitious Long Range Plan

- Selected actions on :
  DMP, Catalogue, AAI, data plateforms
  to speed up transition to Open Science
- Still a lot to do :
  - Improved auxiliary data standardization and collection
  - Coordinated Data-Lakes,
  - Analysis Platforms, Containers technologies for standardized analysis, ...

Contribution on Open Science to the NUPECC Long Range Plan 2024.

Contact us if you are interested in these initiatives