
CERN

# Report from the International, Inter-disciplinary ADMP W/S

**Held at CERN, 28-30 June 2016**

**Agenda, background, videos, notes:**

https://indico.cern.ch/event/520120/

# Background

- Why this (that) workshop?
  - Grew out of several discussions at the RDA, involving RDA ADMP IG, corridor discussions, and the 2nd RDA EU "Data Science Workshop (also held at CERN)
  - **Long & "painful" gestation period (cf elephants)**
- Why at CERN?
  - I am not sure anyone else volunteered.
    And we care a lot about "Data Management"
    - **N.B. DM and (A)DMP are not the same!**
  - CERN bore all local costs;
    travel & subsistence paid by (physical) attendees
- Who participated? All ADMP co-chairs ++
  - But no-one from EU (perhaps as they were revising H2020 guidelines at the time… Now "FAIR DMP")
- **What was the outcome? We will see…**

# Who participated & how?

- Roughly split between local & remote
  - 20 + 20 (not full time – particularly with time zones)
- Remote participation & presentations worked well!
  - Turning video (Vidyo) on to indicate desire to talk / leave off otherwise
- Good split between funding agencies, tool providers, and even “users” (writers of DMPs)
  - HEP talks on experience with data sharing, re-use, reproducibility – both “now” and “then”

# Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the **Office of Science** for research funding are required to include a **Data Management Plan** (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be **shared and preserved** or explain why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable **validation of results**, or how result could be validated if data are not shared or preserved.

# H2020: Annex 1 (DMP Template)

The DMP should address the points below…

1. Data set reference and name
   - Identifier for the DS to be produced
2. Data set description
   - Description; origin; nature & scale; to whom useful; underpins publication? similar data?
3. Standards and metadata
   - Reference to standards *of the discipline*
4. **Data sharing**
   - How will it be shared? Embargo periods? Mechanisms for dissemination, s/w and other tools for re-use, access open to restricted to groups, where is repository? Type of repository?
5. **Archiving and preservation**
   - Description of procedures, how long will it be preserved? End volume? Costs? How will these be covered?

## Workshop on Active Data Management Plans

- Agenda, talks, videos, conclusions
- Includes more detailed talks about HEP data preservation & Open Data releases



## CMS data levels and open data

- CMS has approved a data preservation, re-use and open access policy, which defines the approach to access to them at various levels:
  - Level 1 - Open access publication and additional numerical data
  - Level 2 - Simplified data for outreach and education
  - Level 3 - Reconstructed data and the software to analyze them
  - Level 4 - Raw data, and the software to reconstruct and analyze them.

### CMS Open Data

- CMS continues publishing and promoting levels 1 & 2.
- CMS data releases at level 3 - reconstructed data:
  - November 2014: 28 TB of 2010 collision data
  - April 2016: $>$ 100 TB of 2011 collision data and $>$ 200 TB of simulated data.

▸ CMS data preservation, re-use and open access policy

# Success metrics

- Change the way that funders expect DMPs to be produced
- DMPs – will be able to be created, evaluated, monitored – to enable data to be exploited immediately and over the long term
- Support EU Data Principles
  - Discoverable – readily found to exist by online search
  - Accessible– when discovered they can be interrogated
  - Intelligible– they can be understood
  - Assessable– the reliability of their source can be evaluated
  - Useable– they can be re-used
    - Interoperable (from FAIR)

# Aims of workshop

- Discuss ideas of requirements
  - Guided by success metrics
- **Understand how DMPs can help identify synergies across projects and disciplines**
- Separate group discussions on various aspects of DMPs
- Agree overall concepts
- Detailed next steps
  - Individual responsibilities
  - Schedule
  - Co-ordination

# Actions(?) & Next Steps

- **Future workshop(s) on "extended ESFRIs" searching for synergies using DMPs (JS)**
  - Foreseen for Spring / Summer 2017
  - If successful, repeat as long as useful!
- **Work foreseen in the context of CCSDS (DG)**
  - http://cwe.ccsds.org/moims/default.aspx#_MOIMSDAI to join mailing list
  - These have links to the drafts of the document:
  - “Information Preparation to Enable Long Term Use”.
    - See also wiki http://wiki.oais.info/bin/view/

- **Work foreseen through RDA IG (HG, KA, DG)**
  - iPRES16 Workshop October 36
  - Contribute to the RDA ADMP emails & wiki
  - Collect Use Cases
  - Contacts with funders to understand their needs e.g. for monitoring and compliance
  - Collect information and plans about tools
  - Encodings for exporting ADMPs
  - Make a series of good DMPs public
- **Other (ADMP) workshops to follow-up on key issues**
  - **Do we need multi-day w/s to complement those at RDA, co-located with iPRES, iDCC, … ?**

➢ **Bottom line: a lot of willingness to drive forward, synchronise tools, understand benefits (and drawbacks)**

# *Stakeholder Requirements*

- **[JS1]** Requirements from FAs are not static: compare the H2020 DMP guidelines from 2015 with the new ones from July 2016 (with IMHO too much emphasis on FAIR, which might be a nice concept but do we really understand how to implement the necessary services?)

- **[JS2]** International projects have to deal with requirements from multiple FAs. The ATLAS experiment at CERN has members from around 50 countries. Having to maintain DMPs for each country (for each experiment at CERN) when requirements are still changing is a problem (and not really an opportunity).

- **[JS3]** Do we have the necessary structures / channels in place for a meaningful discussion between the various parties on these (and other) points?

**Open discussion** about the **where** we are, **what** recent requirements are and **how** we could achieve ADMP