International and Interdisciplinary workshop on Data Management Plans

Website incl. presentations: https://indico.cern.ch/event/520120/overview

Tweets: https://twitter.com/hashtag/ActiveDMPs?src=hash

- Local participants (registrants): https://indico.cern.ch/event/520120/registrations/participants

Participants

In person: 18 participants

First Name	Last Name	Affiliation	Country
Kevin	Ashley	Digital Curation Centre, University of Edinburgh	United Kingdom
Frank Olaf	Berghaus	CERN	Switzerland
Eliane	Blumer	University of Geneva	Switzerland
Xiaoli	Chen	University of Sheffield (GB)	
Sunje	Dallmeier-Tiessen	CERN	Switzerland
David	Giaretta	Giaretta Associates Ltd	United Kingdom
~~Helen~~	~~Glaves~~	~~British Geological Survey~~	~~United Kingdom~~
Andy	Gotz	ESRF	France
Marjan	Grootveld	Data Archiving and Networked Services (DANS)	Netherlands
Rob	Hooft	Dutch Techcenter for Life Sciences and ELIXIR-NL	Netherlands

Marie-Christine	Jacquemot-Perbal	Inist-CNRS	France
Isabelle	PERSEIL	INSERM	France
Amy	Pienta	ICPSR	United States
Raphaël	Rey	EPFL	Switzerland
Geneviève	Romier	CNRS	France
Jamie	Shiers	CERN	Switzerland
Kathryn	Unsworth	ANDS	Australia
Dario	Vianello	EMBL-EBI	United Kingdom

Remote participants:

Session 1 - Tuesday 28 June

- Achim Geiser
- Aida
- April Clyburne-Sherin
- Daniel Mietchen
- Frank Olaf Berghaus
- Jamie Shiers
- Maarten Kooyman(SURFsara)
- Matthew Viljoen (EGI.eu)
- Patricia Knezek
- PC-Room-31-3-004
- Vanessa Acín - IFAE
- VidyoPanorama IT Amphitheatre

Session 2 - Wednesday 29 June morning

Participants: 14

- Aida Palacio
- Angus Whyte DCC
- Daniel Mietchen
- Daniela Docan
- Esther
- Felix Engel
- frank

- Frank Olaf Berghaus
- Jamie Shiers
- Matthew Viljoen (EGI.eu)
- PC-Room-31-3-004
- Recorder
- Vanessa Acín - IFAE
- VidyoPanorama IT Amphitheatre

Session 3 - Wednesday afternoon

Participants: 15

- Aida Palacio
- Andrii
- Angus Whyte
- Daniel Mietchen
- Daniela Docan
- frank
- Jamie Shiers
- Justin Noble
- Mark Leggott
- Matthew Viljoen
- Patricia Knezek
- PC-Room-31-3-004
- Recorder
- Vanessa Acín - IFAE
- VidyoPanorama IT Amphitheatre

Participants: 16

- Angus Whyte
- April Clyburne-Sherin
- Daniel Mietchen
- Daniela Docan
- FE
- frank
- Frank Olaf Berghaus
- Heike Görzig
- Jamie Shiers
- Justin Noble
- Matthew Viljoen
- Patricia Knezek
- PC-Room-31-3-004
- Recorder

- Stephanie Simms
- VidyoPanorama IT Amphitheatre

Day 3 (Thursday 30 June)

Participants: 9

- 31-3-004
- Aida Palacio
- Daniela Docan
- Esther
- Jamie Shiers
- Jamie Shiers
- Recorder
- Vidyo Support - Theo Soulie
- VidyoPanorama IT Amphitheatre

Participants: 10

- 31-3-004
- Daniel Mietchen
- Daniela Docan
- Esther
- FE
- Jamie Shiers
- Jamie Shiers
- Matthew Viljoen
- Recorder
- VidyoPanorama IT Amphitheatre

Requirements for (Active) Data Management Plans

Introduction and Workshop Goals (Giaretta & Glaves)

See slides at
https://indico.cern.ch/event/520120/contributions/2164502/attachments/1299256/1938723/Introduction_and_Workshop_Goals.pptx

- Outcome: Next steps on our DMPs
- RDA ADMPs – are these goals reasonable and can they be improved.

- Are there conflicts between different funders requests:
    - Are we planning to change the requirements by funding agencies
- Jamie: Regarding data sharing. EU are very vague on where the resources come from. Reproducibility is very open ended: what do you want to reproduce, where is the funding?
    - David: Somebody has to provide resources
- David: Success metrics: We are successful if we can tell funders how DMPs should be produced. And we may have specific points to change requirements.
    - Do we want to come out of this how to identify resources.
    - Jamie: there has to be a discussion between funders, service providers, and scientists. This way we can determine what is useful. This dialogue is missing.
        - David: Should encouraging dialogue via DMP goal of the workshop?
        - Patricia: Yes we can only fund so much
    - David: Are resource requirements expected in DMPs by funders?
        - Jamie: Yes in some
        - Patricia: Difficult problem
        - Rob: Cloud coins in the OpenScience cloud – cloud providers provide solution and are awarded funding via OpenSciCloud budget
        - David: We should have more discussion here.
- Jamie: avoid “we all agree about ___”. We often don’t understand what we are talking about here.
- Frank: Encourage clear relation to technology to solve your aims. This translate to funding requirements.
- Round table introductions:
    - Jamie Shiers, CERN, DMPs coming for some time, projects in EU and US not funded because of their DMP, We should get some structure and dialogue based on that structure. Large project synergies – S3 roadmap.
    - Geneviève Romier, CNRS/France, Long tail of science needs DMPs, trying to introduce, wants ideas/trends on what can be done. Multiple projects in the same research goal
    - Kevin Ashley, Digital Curation Centre, Share our vision of data management, and get discussion in the field outside of the curation community, vision: improve research and get the most out of our data

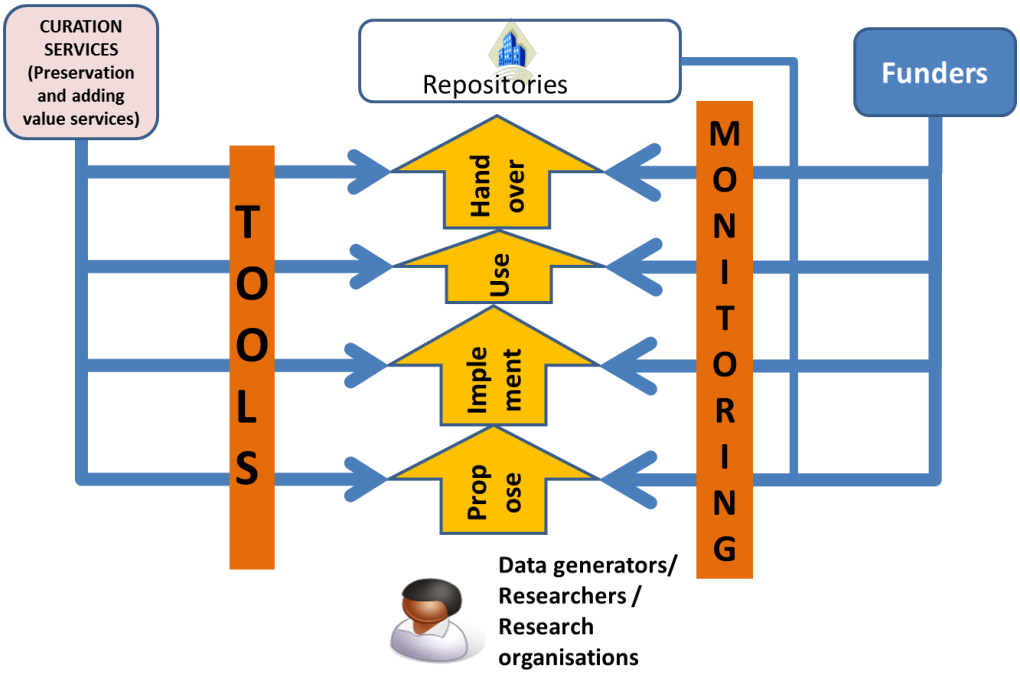
- Xiaoli Chen, CERN, Japanese research far behind, data management highly recommended but not supported. Move involvement in international projects = interest in DM and OA.
- Matthew Viljoen, EGI.eu OpenData platform and Indigo data cloud, wants to contribute pow of end user, and what functionality needs to be provided
- Rob Hooft, ELIXIR, agglomeration of small projects into a large one, View point from the life sciences, DMPs here are useful to teach life scientists on how to deal with their data and how much it can help to do this right. A DMP questionnaire is a checklist to encourage scientists to meet with experts.
- Raphaël Rey, EPFL Library, works on search engine/IT, aims to get an education on DMPs for work in DM for himself and colleagues?
- Marie-Christine Jacquemon-Perbal, IES, trying to organize and aid scientists in writing DMPs from the beginning of the projects. Wants good practices to communicate vie template to scientists. How to deal with physical samples? Practices for file naming and metadata – structure and documentation. DMP should be a tool not an administrative form.
- Marjan Grootveld, DANS (Data Archiving and Networking Services, NL) , convey information on how social scientists work, how they experience data & data management, would like to take home examples/lessons for Dutch funders, In 5 years we should be doing research properly (not talking about DMPs)
- Kati Lassila-Perini, CMS, implementing OpenAccess in CMS, helpful for practices and funding requests from scientists,
- Frank Berghaus, CERN, ...
- Patricia Knezek, NSF, Astronomer in advanced cyber infrastructure, very new herself to DMPs and DM. Contribute questions. How to build a DMP driven by the researchers on what they want to do. How do this without placing undue burden. Want to see lessons learned and other organizations approaches – and how the NSF can use these.
- Daniel Mietchen, NIH, gap between policy and tools, how do we close this gap? DMPs should be facilitator of work and a discovery tool – especially for users and re-users of data. Wants a cross disciplinary and international perspective on what to do next, along with better coordination with other actors in this space.
- Achim Geiser, DESY, data preservation for ZEUS, contribute ZEUS experience,
- April Clyburne-Sherin, Centre for Open Science (US), teaching workshops for researchers on reproducibility, vision to improve quality and impact of research, brings questions, issues and frustrations of researchers. Highlight tools and guidance needed. Find existing tools and practices.

- David Giaretta, PTAB/Giaretta Associates, vision: data should be in the best possible place to be exploited, will contribute cast in stone ideas from years of experience, goal: we should form a core that can move forward DMP and guide their design

- DM/Curation service and tool provider funding is somebody else's problem

The View from ADMP Interest Group (Giaretta & Glaves)

- Slides at https://indico.cern.ch/event/520120/contributions/2164553/attachments/1299259/1938727/The_View_from_ADMP_Interest_Group.pptx
- Gives background – specifically about the Research Data Alliance
- ADMP workgroup in RDA has been inactive after the initial flurry of activity putting ideas on the wiki, because the charter took a long time to be approved. Will come back to this point later.
- A summary of the work done at the start of the ADMP (BoF at that time):
  - There was a view at the meetings that DMPs are currently too short to be useful/inadequate
  - Initial ideas are on a wiki https://rd-alliance.org/groups/active-data-management-plans/wiki/active-data-management-plans-wiki-index.html where an overall picture was drawn up and some potential Working Groups were suggested.



  - Summarising the discussion during the presentation David noted
    - Identify how many challenges are shared between vs unique to projects
    - Bottom up motivation for DMPs will be interesting to see

- Data citations have direct benefits (to data creator)
- DMPs should expose continuity of practices between projects

CERN Experience & Plans: (A)DMP and more (Shiers)

- Re “long term”: CERN has existed since 1954. An upgrade of the HL-LHC has been approved, meaning that LHC data have to be re-usable until (at least) the second half of this century.
- LHC computing started with R&D projects from 1994. “Grid-itus” from 2000 and serious application of the Grid from 2004, with accompanying challenges, which get easily forgotten about.
- The different LHC centers across the world have different roles in terms of generating, analysing and preserving the data, e.g. regarding Tier 0 (at CERN) - Tier 1(11 centers) - Tier 2 (over 200 centers). Thousands of users worldwide, so “talk to your users” isn’t easy.
- “Long-term” also means for projects this scale that you have to plan and stick to it (live with it) for a long time. Looking around to projects of similar scale, e.g. EBI.
- Importance of software in search for the Higgs Boson rarely stressed - first time on 4 July 2012 by CERN Director, who mentioned the software along with the machine itself and the experiments run on it.
- Raw data is at CERN plus a copy spread across Tier1s, but where is the data behind the publications? Shared responsibility!
- Complexity: DMPs required by 21 member states, plus many non-members, plus many experiments.
- DPHEP business case and cost model needed. Common set of uses cases was agreed across major HEP experiments, as well as “value” in terms of publications, PhDs etc.
- Explicit agreement to fund long-term preservation, not data sharing.
- HPC use cases turn out to match very well with the overall DMP requirements and the business case.
- LHC experiments have data policies, which Jamie considers to be kind of extended DMPs.
- CERN (WLCG) is preparing to acquire the ISO 16363 certification as Trusted Digital Repository. Current status: not just WLCG, but also “CERN’s Digital Memory” (photos, videos etc.) [Feedback from the audience: very good to combine this]. (Self-)certify site-wise, with variations for other sites, e.g. because of their different designated communities.
- How does/ should a disaster response plan for IT look like? At CERN, currently a Word doc sent around via email (??!)
- Data preservation and certification of TDRs help address the goals of FAIR & Open data and DMPs. Note that “Open” would not per se open for the general public: could be on demand e.g., also depending on resources. Still, the DMP should state the ambition to share, along with potential limitations.

- Compared to other disciplines, the data *access* costs are relatively high, compared to bit preservation costs.
- Question from the audience: have you tried to formalise data policies into something machine-readable? Could be interesting for statistics and reproducibility. Answer: for ALICE, the goal is that all details that go into a publication are fully captured in a reproducible fashion.
    - I think this may be a reference to the CERN Analysis Preservation project? ATLAS for example collaborating with that and the RECAST Project to capture a machine actionable analysis workflow.

Discussion – Who should benefit from DMPs?

- Daniel Mietchen: the funders should get a better insight into the usage of research data and associated computational and preservational infrastructure, but ideally, DMPs should be used in day-to-day live by researchers and research teams to manage, discover, analyze and otherwise handle data. Other beneficiaries include data centers, funders, educators, students, private project partners, journals that require that reviewers can access data underlying submitted publications.
- Kevin Ashley: there is a huge waste in writing DMPs and not having them reviewed. A DMP in practice is/ should be a second hurdle, after having written the project proposal.
- Patricia Knezek: ideally the DMP is a good secondary quality check, but especially for smaller projects it is a challenge to do *long term* data preservation. Problem is not in the planning.
    - Rob: in some disciplines there are already copies of data, so 'contact a good repository and don't try to organise preservation yourself' would be good advice.
- David: the DMP can help the various parties involved to plan their part themselves. Try to integrate all roles.
- Matthew Viljoen: the users themselves, to account for what they did.
- Kevin: having a larger infrastructure, including the long-term data centers, helps researchers with planning and with selecting which data to keep and which data to let go.
- Kevin: Users can not be expected to foresee all potential commercial applications of their data.
    - Rob: Funders do not necessarily want data to be as open as possible, they want the best societal value of the data. Sometimes that requires data to stay closed (for a while?), e.g. while applying for a patent on a drug.
    - Marjan: Sometimes projects find out half way that involved companies want to keep data closed.
        - Data management planning could expose this beforehand
        - Active data management planning can deal with the change of plans.

- Rob: users are often also producers of data, we all benefit from the standards that each of us adheres to. Keeping as much data as we can is a good practice: we've seen examples of unexpected value of data much later on.
- Discussion going back and forth now, harder to keep track of. Some themes:
    - Involving re-users early on?
    - Coupling between data and metadata, and sometimes only keeping the metadata [+ other context information - note from the note taker] is sufficient.
    - Findability in the context of the FAIR principles http://www.nature.com/articles/sdata201618 - example of images taken to analyze the spine might be useful for analyzing the aorta, so describe what you have in terms of data, not (solely) what you're using it for
        - Another example: a project that set out to study biofluorescent corals did eventually provide data on biofluorescent fish, the first reports ever of biofluorescence in vertebrates
        - Jamie: it is not clear that findability is always desired, e.g. the fact that someone has undergone a test for STD's may be as sensitive as the outcome of that test. Rob: there are developments ("Beacons") that are designed to deal with this kind of sensitivity.
- Marie-Christine: the librarian can help to document the data - to be involved in the planning stage.
- Daniel: could we already agree on something to explore doing together, e.g. a common use case?

Experience with Data Management Plans

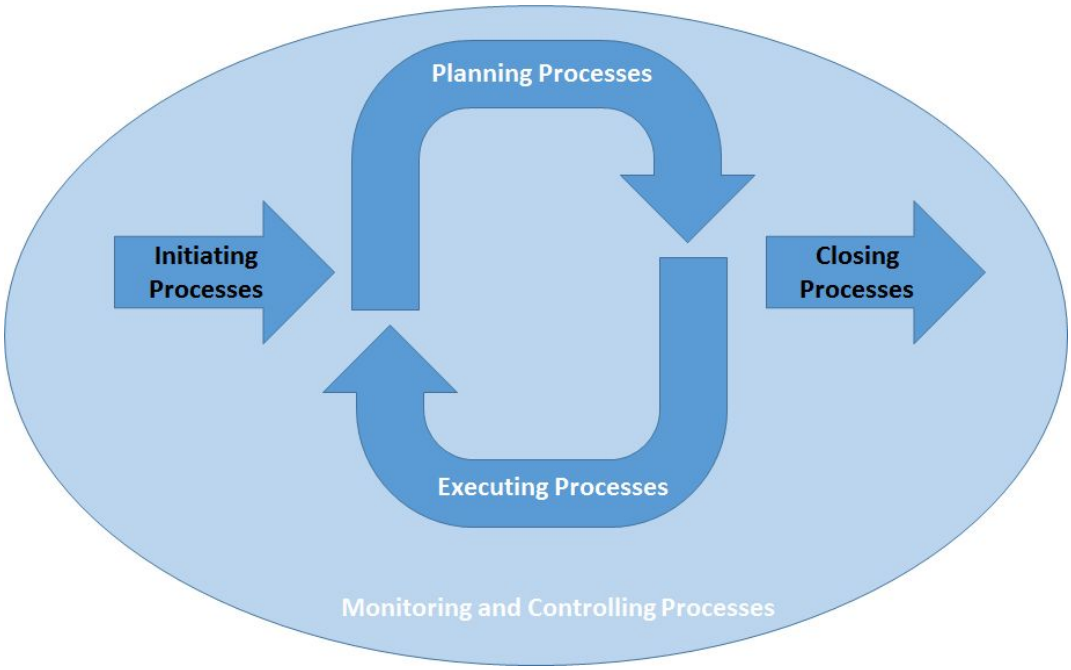
- Notes

Active Data Management in Space (Giaretta)

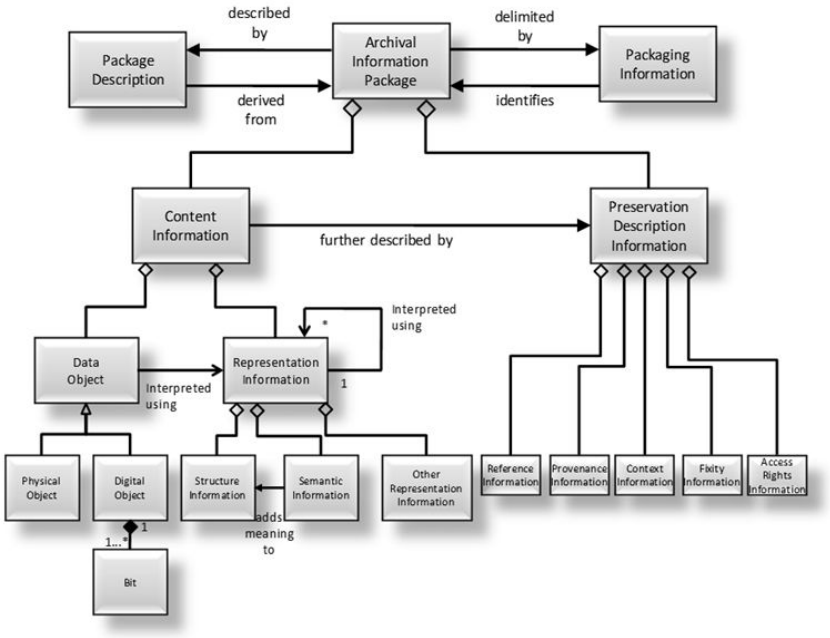
See slides at https://indico.cern.ch/event/520120/contributions/2171073/attachments/1299260/1938729/Active_Data_Management_in_Space.pptx

- CCSDS/ISO, the group that produced the OAIS model, is also working on a standard which is relevant to ADMP.
- This document should complement the RDA-ADMP work and was started because of the delays in the RDA and also because the RDA publication process is only just starting whereas ISO standardisation is a well developed and ISO standards well known.
- The document will pass through both the CCSDS and the ISO review processes and if successful will become a full ISO standard.

- The document has a working title: “Information Preparation to Enable Long Term Use” and embodies several of the ideas from the ADMP wiki. It should be understood that this document is focussed on ensuring that data which is created/collected can be understood /used /exploited now and into the future by ensuring the correct data about that data (*i.e. “metadata”) is captured. ADMPs will undoubtedly cover other topics.
- David noted that the term “metadata” tends to be overused and causes misunderstandings. It is better to be specific about what you are describing instead of using the blanket term metadata. OAIS provides many specific terms.
- The standard being drafted uses concepts from OAIS and the “Project Management Body of Knowledge” (PMBOK, see http://www.pmi.org/PMBOK-Guide-and-Standards.aspx). The original ADMP diagram showed 4 stages of a project which is creating/collecting the data, however is is rather a crude way to describe activities which may range from the small to the extremely large and complex. PMBOK uses a much more widely applicable concept of “Process Groups” rather than stages.



- PMBOK then proposes a number of Knowledge Areas and then produces a table which summarises which Knowledge Area should be active in each of the Process Groups.
- Rather than define Knowledge Areas, CCSDS is using as a basis the OAIS concepts defined in the Archival Information Package (part of the OAIS Information Model) - because this identifies the pieces of information needed for long term preservation of the data being preserved.



- It should be recognised that OAIS judges preservation based on continued understandability/ usability, therefore its concepts can be used if one is interested in broader ideas of usability.
- In addition to the Information Model there are other pieces of metadata which must be collected.
- It should also be understood that in general plans for managing data will involve more immediate concerns such as how the project handles the data while it is being collected.
- Bringing the PMBOK Process Groups and the OAIS Information Model together the document has a table of information which should be collected - and improved - through the process groups. For example the Representation Information which must be captured will be quite poorly understood when initiating, but as further planning and executing occurs, the Representation Information should be much better known and at closing it should be essentially complete.

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Representation Information	Choice of data format	Rough idea	Increasingly detailed	Becoming complete	Complete
	Format definitions and	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating

formal descriptions				
Semantics of the data elements	Rough idea	Increasingly detailed	Becoming complete	Almost complete
Data dictionaries and other semantics	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
Information Model	Rough idea	Increasingly detailed	Becoming complete	Complete
Other Data Documentation	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
Applicable standards	Rough idea	Increasingly detailed	Becoming complete	Complete
Hardware and Software Dependencies	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
Other software which may be used on the data		Increasingly detailed	Increasingly detailed	Growing
Calibration and system test tools and system test data that will be delivered.	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
Relationships between data items	Rough idea	Increasingly detailed	Complete	Complete

- The presentation has snapshots of other parts of the table.
- We should set a **follow-up workshop** on Thursday. David will share the details of CCSDS for further discussion.

Meanwhile, back on planet Earth (Glaves)

- Sadly speaker could not make it

ADMP in Life Sciences (Hooft)

- Life science is becoming a data science
    - Life sciences = Food, disease, ecology, bacteria, applications in biotechnology, etc.
    - Biologists/Medical scientists are not computer experts, or interested in becoming those: bioinformatics is the intermediary
    - Problem: no trust in existing tools. They are not adopted, everybody writes their own tools. My problem is unique, but my solution is generic... https://en.wikipedia.org/wiki/Dunning%E2%80%93Kruger_effect
        - Exists on the compute side as well: example computing and storage are each complex and important aspects of data management
- Technology evolves - eg sequencing has become much more efficient since 2005 - and produces much more data.
    - Increased data volume goes hand in hand with increased computing needs
        - Measurement and analysis tools and techniques change
    - Eg: one human genome can be measured in ~48h the compute now takes ~300hrs - modern technique is very different from that used in 2005.
- Very varied data: O(10^6) life scientists with very different projects
    - Small projects sometimes underestimate costs of data management by O(100)
    - Genetics, proteins, chemical analysis, people measuring single compound trying to make correlations. Data collected by doctors in hospitals about patients is 80% text and only 20% structured data.
    - Data munging is a huge problem: this can be reduced by the Interoperability in FAIR.
        - Future use of life science data requires wide interoperability, e.g. Wheat grown in Australia has to handle temperature increase: intersection of climate, genome, and wheat (geolocation) data
    - During your project you can use whatever standards for data you like. Data management plan has to account for converting/storing the data in a way that other people can understand. Comment from the audience: it certainly doesn't hurt to advise researchers to use the 'proper' terminology from the start.
    - Data management is different from and guided by the data management plan.
    - "Have you thought about how others might use your data?"
    - Data Management Plans need to be active because "...plans are useless, but planning is indispensable" - Dwight Eisenhower
- Slide on "cloud coins" (that came up yesterday): Funding agency and DMP: funding agency does only need to know that there is a DMP and that it has been approved by some authority (similar to the procedures already widely in use for reviewing the ethical aspects of a project)
    - Cloud Coins is in an idea/design stage currently

    - Providers need to be self certified to some standards
    - Cloud data service providers self certify. This is not unheard of as procedure: The CE mark on electronics products (for example) is not enforced/checked until a user complains

ESRF Data Policy (Götz)

- There are about 30-40 synchrotrons in the world. Amongst the things they do are protein structures, e.g. much of the data in the Protein Data Bank. Of these synchrotrons, most do not have an equivalent of a DMP.
- "We don't write the publications, .. we just produce the data", and "if data is your core product, you ought to curate it properly."
- When the metadata is well managed, it will contain much information
- Adopting new technologies (tape in this case) changes cost and capabilities of the archive
    - Our resources are limited, so we just consume off the shelf resources
- For cost reasons, data were deleted after 50 days. Now, the infrastructure is there for synchrotrons to provide the data on a much longer time scale.
- Current DMP is derived from PaNData
- CC0 seems the sensible choice: citation and reference should be by the honour system. Otherwise we would still have to cite Gauss every time we use a Normal Distribution.

Data management aspects in the social sciences (Grootveld)

- DANS: Funded since the 1950s to keep scientific data around for the long term
- More and more disciplines adopted the service over time
- Matthew effect: Data that is open and accessible is downloaded/re-used more frequently
- Large parts of the data not used for long periods of time - likewise in other long-term data archives
- Actively trying to encourage reuse; cultural barriers to that
- Short (2 page) DMP is a good idea. Makes the DMP clear and easily readable
- Dublin core is good for finding data- but not for learning how to use it
    - Data Documentation Initiative is adopted to answer this lack
    - Very useful for research process and linking between studies (used in humanities in surveys and interviews)
- Submerge yourself into the exercise on how to best organize your data
- Echo: Planning is important, plans not so much.

- Don't just think of data management planning, but broaden it in scope to include planning of software, specimens and possibly other technical/ logistic aspects of the project
- Make intelligent use of versioning
- Software and its preservation should require planning as well, see http://software.ac.uk/software-management-plans

What are the drawbacks of current DMPs? (Giaretta & Glaves)

- Is there a value to the plan, not just to planning?
- DMPs should be able to help addressing legal and ethical access rules and regulation.
- Access rights statement could be made machine readable, but there are still issues covered by DMPs which are simply human communication
- Open Rights Language can be used to create machine readable DMPs (is it ODRL?)
- Careful: If you don't write the plan, but let a machine produce it, what is the benefit?
- Some private/non-government funders now require DMPs

Planning versus the plan; Balancing benefits (Ashley)

- There has been some discussion in the DCC community whether or not to add good practices in the DMP writing guidance: researchers should not just copy them but think for themselves < > as service providers we're used to make their life easy. [Question from note taker: why does this concern pop up more frequently with DMP writing than with research proposal writing?]

Research Data Canada / Canadian viewpoint (Leggott)

- RDC is the link between researchers and funders
- Canadian public funders have no requirements for DMP, but it is suggested in "guiding principles"
- https://portagenetwork.ca/ is a customised version of DCC's DMPonline. Canadian research funders actively promote use of RDMPs by researchers.
- CASRAI is very interested in making DMPs easier and more standard, e.g. using vocabs/taxonomies for data entry
- DRUPAL is an easy technology to implement auto-generated DMPs from questionnaires - but QML with DDI is a better machine readable language.

NSF Viewpoint (Knezek)

- The viewpoint from a funding agency
- The goal of the agency is to promote science (and thereby health, wealth, etc.) - promote education and workforce in STEM
- NSF sees the DMP as a burden to researchers, and aims to minimize the burden
- Data sharing involves many aspects - even administration - that needs to be considered
- Physical security of the data is in fact important - are these measures consistent with law?
- The NSF subject directors are looking into how the enforcement of DMPs has changed data access patterns by researchers

Making DMPs machine-actionable and public (Ashley for Simms and Mietchen)

- Slides at https://indico.cern.ch/event/520120/contributions/2177900/attachments/1299358/1938982/DMPs_-_actionable_and_public.pdf
- Multiple stakeholders in data management - keep in mind the benefits of DMP for these different groups
- Publishers' role to managing data is minor, they should be included after workflows fitting other Stakeholders are defined
- DCC has identified the themes covered in DMPonline, for instance for tagging DMP questions. The theme list is currently offered to administrators for customising their instance of the tool. Will be reduced to less items, implemented for DMP templates and then the tagging can be tested within the community.
- Discovery is a use case of machine-actionable DMPs, e.g. watching out for new data in a particular area or automatically searching DMPs for who is working on the Zika virus outbreak.
- Funder use case, e.g. compliance checking of data deposit in named repositories by checking the DOI.
- Ambition: DMPonline API to *create* a plan. Would be nice to link up with Current Research Info Systems (CRIS), but experiences with them are mixed.
- There is a funding problem here: who is going to fund data management after the experiment/project completes? [In the context of the OpenAIRE project (and probably elsewhere as well) the National Contact Persons have brought this up with the EC. “RDM costs are eligible for funding”, but only within the duration of the grant agreement... - note taker]
- What are actions machines should be able to take in response to a DMP:
    - Ping a service
    - Adopt a storage policy (IRODs)
    - Gather statistical power, for e.g. in psychology

- Find out who is doing what
- Start conversation between different research stakeholders?

What is holding DMPs back? (Giaretta & Glaves)

- We have talked very much about Findable and Accessible: what about interoperability and reuse
- The benefits need to be more clearly stated.
  - We should monitor improvements in data use due to DMPs as noted by Patricia in “NSF Viewpoint”
- What are the tools supporting/implementing DMPs
  - These are communicated in the DMP of the experiment/project
  - When these are published it makes the decision making process open, and allows discussion and improvement
- Funder mandates increase as we make data management practices possible
- “OPEN” data is a dangerous misnomer that came into being because of the parallel development from closed publications to open publications. But we are not talking about opening up data, but about changing from “no data at all” to “described data”. Researchers dealing with privacy sensitive data immediately close the book as soon as “open” data comes up: they incorrectly think it does not apply to them.

Experiences, Challenges and Issues

CMS Data Policy and Open Access Release Experiences (Lassila-Perini)

- Slides at https://indico.cern.ch/event/520120/contributions/2165275/attachments/1301142/1942497/CMSOpenDataDMPJune2016.pdf
- Open Data Portal http://opendata.cern.ch
- CMS (Compact Muon Solenoid) is an experiment running at CERN,the i.e CERN is the service provider.
- Data levels identify the type of data (raw, derived, etc.) plus the software needed to use them
- Kati notes that CMS is very good for “immediate” metadata (mostly Provenance), but had not been so good at the “understandability/usability” (Representation Information)

- Some information is so obvious that it is easily forgotten that it should be documented. A policy to make everything open actually helps with this because it becomes obvious that outsiders will need this obvious information too.
- Examples are a good form of documentation for future re-use.
- OpenData allows testing whether preservation information and data management is sufficient by an outside community
    - OpenData also identifies unexpected user communities
- High level validation of the data by testing statistical distributions
- Note on FAIR: search/find is not the same as discovery: http://sethgodin.typepad.com/seths_blog/2014/04/search-vs-discovery.html. This is a greyscale: start with efforts to make data available to close colleagues and work your way out to larger discoverability. The ideal is that scientists that could use the data would even be able to find it without knowing that it exists.
- Strong competition about resources: for preservation or for data taking, operation and new analysis. Data storage in the the context of CERN is very small, but still needs to be acknowledged.

Mapping DMPs to a past experiment – the ZEUS experience (Geiser)

- Had a data management plan, but did not call it that.
- Unique data of proton-lepton collisions, to find the structure of the proton. Useful for analysis of the LHC data.
- Example of DMP driven from the bottom up. Finding manpower for the long term management is an issue that needs addressing.
- Data must include data, metadata, software, plus human info: knowledge and usability.
- The data management is distributed over a few sites & compute centers.
- Maintenance of sw, simulation and analysis framework required 4 fte/year + IT > not sustainable long term
- Data is very mappable to other physics experiments, but the exact naming and data formats is different. There is no funding to bring this all to the same page. (That would be “interoperability”)
    - Careful: This will make data more readily understandable - but not necessarily interoperable. The scope of the interoperability needs to be stated in the DMP?
    - Interoperability is a grey-scale. Having a common ontology to refer to different manifestations of effects and particles will help already over using the term “electron” or “e” without explaining exactly which manifestation you mean.
- Direct scientific output for Zeus based on the preservation efforts are about 10% of the total. And the cost of the effort is much less than 1% of the total cost of the project. This is a low estimate of the benefit. Jamie comments that the recent papers could be better papers because understanding of the experiments has improved. In Zeus it “at least did not decrease”.

- When you stop data preservation it stops being usable. Continuous efforts are required to keep data usable.
    - In life sciences, where there are many people making similar measurements on different samples and for different reasons but using comparable equipment, the preservation efforts can be centralized.
- The ideal would be to keep the data uncoupled to a particular data analysis technology (possibly with the exception where the technology can be guaranteed). Technologies like “docker” containers can help keep software running in exactly the same virtual environments, alternatively “Ansible” scripts can be maintained to provision new base systems to run the same software. “Docker” is quite a good marriage to data preservation efforts.
- Usability of information requires a data librarian that can help users around in the data library.

Action and Next Steps

How do we get things moving? (Giaretta, Glaves & Shiers)

- Future workshop(s) on "extended ESFRIs" searching for synergies using DMPs
    - Could invite individually/gradually or all together
        - Who's going to http://www.digitalinfrastructures.eu/ in September? Kevin will see if the DCC timeslot could also be used for this.
    - Identify synergies / common requirements/ tools
    - Annual meetings?
    - Include "long tail" representatives?
    - **Jamie** will be taking action, don't be surprised if such a meeting is organised.
- Work foreseen in the context of CCSDS (consultative committee for space data systems)
    - See http://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI to join mailing list and see email archives. 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/
    - Document needs review before finalisation - please read it and email comments and also join the discussion (Tuesdays at 3pm UK time - Webex details emailed to the list before the meeting)
        - Could group the Additional Information topics using FAIR and identify which (the ("I" and "R") will require lots of details which can only be added later.
- Work foreseen through RDA IG
    - iPRES16 Workshop October 3-6 in Bern: The aim of this workshop is to build a roadmap of a common view on recent ADMP related activities to determine:
        - /where/ we are in the ADMP realisation
        - /what /the recent problems are that prevent the ADMP realisation
        - /how /ADMP requirements could be realised by means of computer-actionable policies
        - /which/ tools are required to implement such policies
        - NB: submission deadline is July 31
    - Contribute to the RDA-ADMP emails and wiki https://rd-alliance.org/groups/active-data-management-plans/wiki/active-data-management-plans-wiki-index.html
    - Collect Use Cases e.g.
        - Data capture
        - Using data
            - Research/interop
            - Reproducibility
            - Training

                - Sustainability
            - Peer review
            - Sharing plans
            - FAIR support
            - Hierarchy of plans
            - Data discovery
                - Proposals of potential data creation
            - (Machine) Actionable DMPs
        - Contacts with funders to understand their needs e.g. for monitoring and compliance
        - Collect information and plans about tools
            - ...
        - Encodings for exporting ADMPs - e.g.
            - Think about DDI related s/w ? Need to see the s/w behind it - **Frank**
            - Ontology?
            - APIs? E.g. DMP web tools - after collecting appropriate Use Cases
            - Things to support interoperability of plans with themselves and CRIS
            - Note that we should try to re-use existing “standards”/mechanisms
- Action to create statement about DMP requirements for circulation to funders and data creators: **Kevin, David, ...**
- Identify benefits to data creators / funders / society
- Other ADMP workshops to follow-up on key issues: when, where, who?
    - Physical meetings with RDA may be difficult, so we should consider Virtual meetings. We will pencil in a schedule every month (aim for 10 a year).
    - ACTION **David** to send Doodle Poll to choose time/date.
- ... to be continued by everyone who likes to :-)
- Make a series of good DMPs public
- Other actions and schedule
- Advertise finalised version of Google doc (as PDF) next week
    - Action: **Jamie** to publish
    - Action: Frank will lock document (at end of next week: 2016.07.08)
    - Action: everyone to help clean up, add references and make more readable, plus integrate their own notes.
- Seek funding to carry on this work when we have a set of clearly identified tasks