Research Data Management

and the Nikhef Policy Draft

Nikhef

WAR meeting

David Groep
PDP

07 December 2017

davidg@nikhef.nl

Research Data Management

Towards long-term data preservation and re-use for research at Nikhef

Introduction

- *Research Data Management & 'FAIR'*
- *NWO Policy Framework Requirements*
- *Our own Data Preservation tradition*

Discussion

- *Nikhef Data Management Policy terms*
- *Principles*
- *Practice Guidance*

Why Research Data Management?

RDM in support of the proper research process

- implementing the VSNU/NWO Code of Conduct* – 'verifiability' criterion
- "Open Science" & "Open Access" policy aims
- conventional re-use for our own purposes ("where was that dataset again??")

but there may be subsidiary aims for RDM

- re-use in different domain: *complex here, but not uncommon for societal challenges*
- citizen science: *e.g. "Hanny's Voorwerp" [J094103.80+344334.2] in astronomy*
- reuse of results for industry, civil society and general public – it's in our mission ☺

http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/Code_wetenschapsbeoefening_2004_(2012).pdf

The “FAIR” Principles

… since ‘open access’ ‘open data’ is basis for (funding) agenda OCW/NWO/KNAW/EC …

pick a name for the effort that is ‘hard to disagree with’ …

FAIR

- Findable - *data & metadata are easy to find by humans & computers*
- Accessible – *standard download method & ‘protocols’ for access explicit*
- Interoperable - *they can be automatically combined with other data*
- Reusable - *sufficiently well described to be replicated or combined*

but this is easier said than done, and implementation domain specific

Wilkinson et al. **The FAIR Guiding Principles for scientific data management and stewardship** *doi:10.1038/sdata.2016.18*

FAIR – a ‘slight’ LS bias

- **To be Findable:**
- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

- **To be Accessible:**
- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

FAIR – more microstatements

- **To be Interoperable:**
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. *'let's really think that you can replace researchers with automated systems'*
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

- **To be Reusable:**
- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

The NWO DM Protocol

- The “FAIR” principles inspired the **NWO DM Protocol**, *and the H2020 Open Data guidance, and ERC requirements*
- Based on the DM Protocol, there is also an ***NWO Institute Data Management Policy Framework***
- Endorsed August 2016, now needs to be implemented soon(ish)

For ‘us’ RDM is not new

last big e^+e^- machine closed years ago,
and *e.g.* LEP, or HERA, data is not re-measurable today …

DPHEP (2008 – *ICFA-DPHEP Study Group*, 2013+ *DPHEP MoU*)

- preserve experimental data *and* its software environment
- study group, repository development (Zenodo), software curation

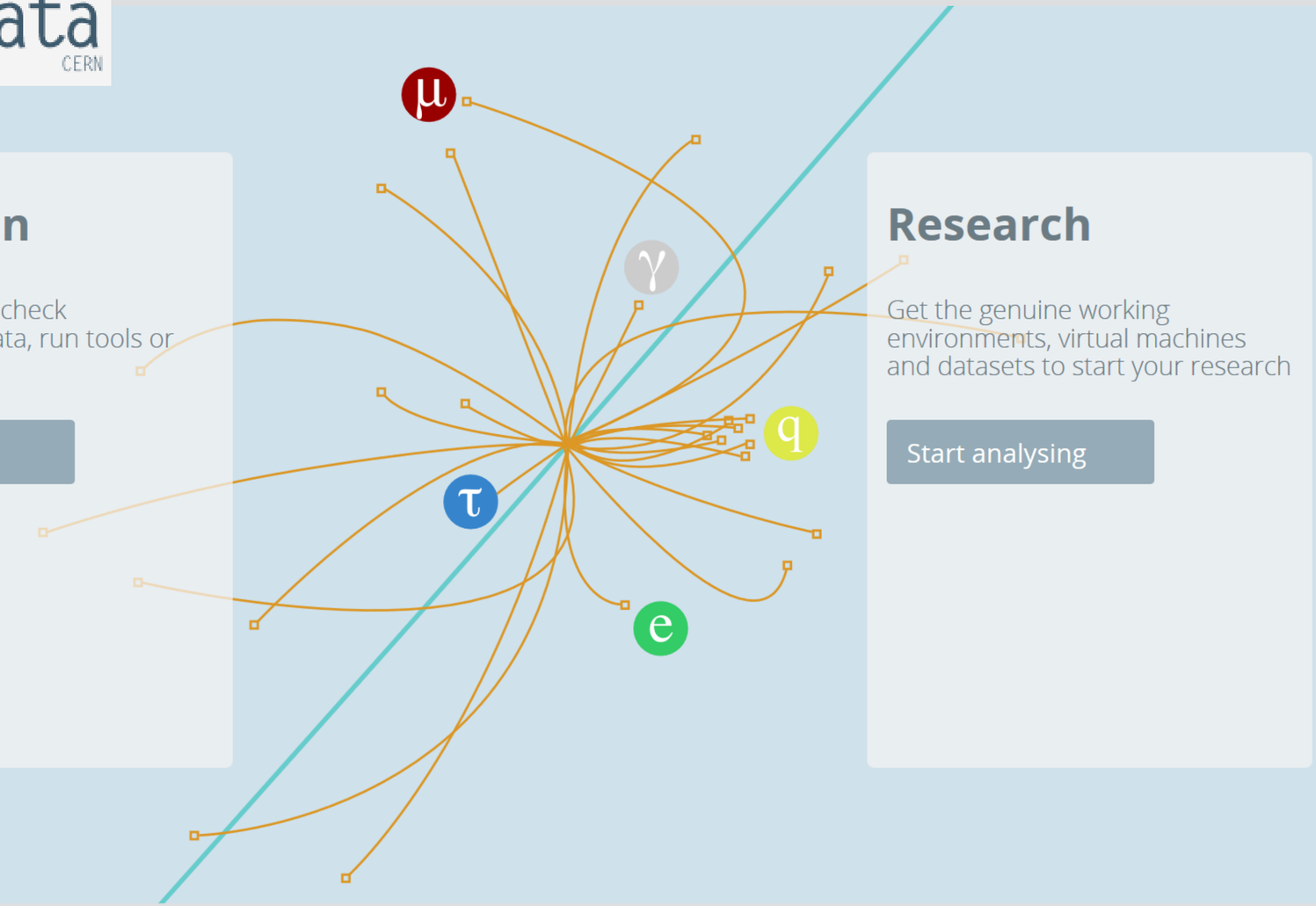
opendata.cern.ch

opendata CERN

Education

Visualise events, check reconstructed data, run tools or build your own!

Start learning



Research

Get the genuine working environments, virtual machines and datasets to start your research

Start analysing

CMS

To analyse CMS data, a Virtual Machine with the CMS analysis environment is provided. The data can be accessed directly through the VM. In the primary datasets, no selection nor identification criteria have been applied. The 2011 data release includes simulated Monte Carlo datasets, but no simulated datasets are provided for the 2010 release.

Explore CMS ›

ALICE

According to the ALICE data preservation strategy, reconstructed data and Monte Carlo data as well as the analysis software and documentation needed to process them will be made available on a time scale of 5 years (for 10% of the data). Thus, the first release of ALICE research data will happen in 2018.

ATLAS EXPERIMENT

According to the ATLAS Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

LHCb

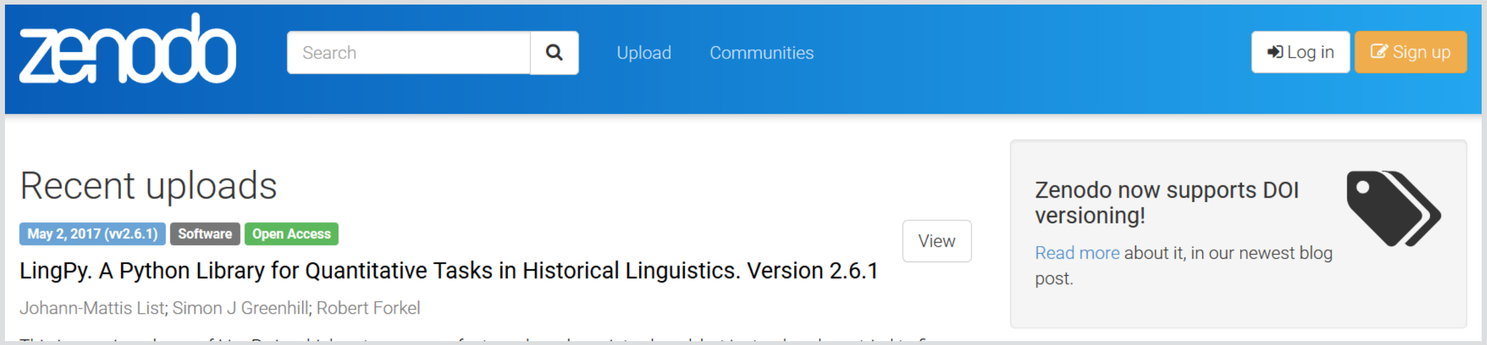
According to the LHCb External Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

We have tradition here!

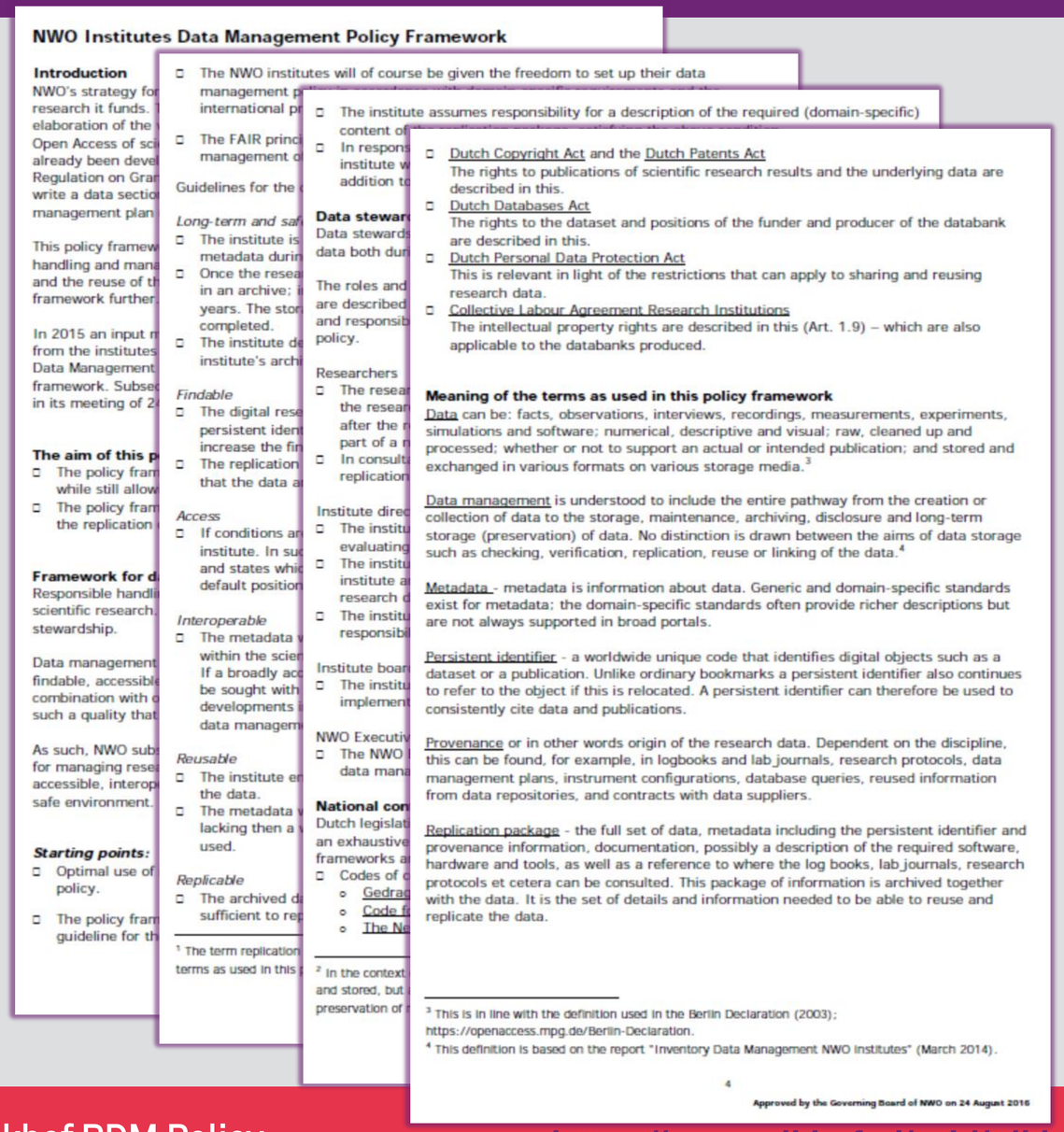
- DPHEP
- HEPForge
- HepData
- CERN OC3
- INSPIRE-HEP ('SPIRES')

and we 'spread the word' through general-purpose services

- **Zenodo**



NWO-I DM Policy Framework



NWO Institutes Data Management Policy Framework

Introduction

NWO's strategy for
research it funds.
elaboration of the
Open Access of sci
already been devel
Regulation on Gran
write a data sectio
management plan

This policy framew
handling and mana
and the reuse of th
framework further

In 2015 an input
from the institutes
Data Management
framework. Subse
in its meeting of 2

The aim of this p

- The policy fram
  while still allow
- The policy fram
  the replication

Framework for d

Responsible handli
scientific research,
stewardship.

Data management
findable, accessibl
combination with
such a quality that

As such, NWO sub
for managing rese
accessible, interop
safe environment.

Starting points:

- Optimal use of
  policy.
- The policy fram
  guideline for th

- The NWO institutes will of course be given the freedom to set up their data management p
  international p
- The FAIR princi
  management

Guidelines for the

Long-term and saf

- The institute is
  metadata durin
- Once the resea
  in an archive; i
  years. The stor
  completed.
- The institute de
  institute's archi

Findable

- The digital rese
  persistent ident
  increase the fin
- The replication
  that the data a

Access

- If conditions ar
  institute. In suc
  and states whic
  default position

Interoperable

- The metadata
  within the scie
  If a broadly acc
  be sought with
  developments i
  data managem

Reusable

- The institute en
  the data.
- The metadata
  lacking then a
  used.

Replicable

- The archived d
  sufficient to rep

[1] The term replication
terms as used in this

- The institute assumes responsibility for a description of the required (domain-specific) content of
- In respons
  institute w
  addition t

Data stewar

Data stewards
data both dur

The roles and
are described
and responsib
policy.

Researchers

- The resea
  the resear
  after the r
  part of a r
- In consult
  replication

Institute direc

- The institu
  evaluating
- The institu
  institute a
  research d
- The institu
  responsibl

Institute boar

- The institu
  implement

NWO Executiv

- The NWO
  data mana

National con

Dutch legislat
an exhaustive
frameworks a

- Codes of c
  - Gedrag
  - Code f
  - The Ne

[2] In the context
and stored, but
preservation of

- Dutch Copyright Act and the Dutch Patents Act
  The rights to publications of scientific research results and the underlying data are described in this.
- Dutch Databases Act
  The rights to the dataset and positions of the funder and producer of the databank are described in this.
- Dutch Personal Data Protection Act
  This is relevant in light of the restrictions that can apply to sharing and reusing research data.
- Collective Labour Agreement Research Institutions
  The intellectual property rights are described in this (Art. 1.9) – which are also applicable to the databanks produced.

Meaning of the terms as used in this policy framework

Data can be: facts, observations, interviews, recordings, measurements, experiments, simulations and software; numerical, descriptive and visual; raw, cleaned up and processed; whether or not to support an actual or intended publication; and stored and exchanged in various formats on various storage media.[3]

Data management is understood to include the entire pathway from the creation or collection of data to the storage, maintenance, archiving, disclosure and long-term storage (preservation) of data. No distinction is drawn between the aims of data storage such as checking, verification, replication, reuse or linking of the data.[4]

Metadata - metadata is information about data. Generic and domain-specific standards exist for metadata; the domain-specific standards often provide richer descriptions but are not always supported in broad portals.

Persistent identifier - a worldwide unique code that identifies digital objects such as a dataset or a publication. Unlike ordinary bookmarks a persistent identifier also continues to refer to the object if this is relocated. A persistent identifier can therefore be used to consistently cite data and publications.

Provenance or in other words origin of the research data. Dependent on the discipline, this can be found, for example, in logbooks and lab journals, research protocols, data management plans, instrument configurations, database queries, reused information from data repositories, and contracts with data suppliers.

Replication package - the full set of data, metadata including the persistent identifier and provenance information, documentation, possibly a description of the required software, hardware and tools, as well as a reference to where the log books, lab journals, research protocols et cetera can be consulted. This package of information is archived together with the data. It is the set of details and information needed to be able to reuse and replicate the data.

[3] This is in line with the definition used in the Berlin Declaration (2003): https://openaccess.mpg.de/Berlin-Declaration.
[4] This definition is based on the report "Inventory Data Management NWO Institutes" (March 2014).

4

Approved by the Governing Board of NWO on 24 August 2016

Implement (by ~1 Sep 2017) the NWO-I DM Policy Framework

- only 4 pages (good ….)
- binding (unlike NWO Protocol)
- Institute Data Managers were consulted beforehand
- Leaves us freedom of implementation, but does require specific elements like the Replication Package

https://www.nikhef.nl/grid/nikhef/dmp/NWO-Institutes-Data-Management-Policy-Framework.pdf

Nikhef RDMP ToR

- must implement the NWO-I DM Framework and NWO Protocol
- not undermine any past or future data management customs already in use in the sub-atomic physics domains (HENP & others)
- leverage as much as possible international efforts (such as DPHEP)
- minimal impact on PhD students and their time
- limit impact also for staff members as much as possible
- not incur unnecessary costs or liabilities

The Document ...

Nikhef Research Data Management Policy

The Dutch National Institute for Sub-atomic Physics Nikhef, via its mission and through the programmes, projects, and collaborations that it operates and subscribes to, is a significant producer of scientific research data, and transfer of this knowledge to third parties, i.e., industry, civil society and general public, is an integral part of Nikhef's mission. Nikhef is committed to ensuring careful management and optimal exploitation of the research data, both in the short term and the long term, in alignment with the principles on data management of NWO, and in accordance with this Policy.

Scope

This Policy applies to all research data that are relevant for re-use and produced as a result of Nikhef *Research Activities*, i.e.,

- all approved granted research programmes and granted research projects, and
- research projects so designated and approved by the Nikhef director, and
- any activity that results in *Published* data as per the General Principles.

This Policy shall apply without prejudice to provisions set forth in more specific agreements between Nikhef and any third party, which in all cases take precedence.

This Policy does not apply to

- data resulting from or relating to work carried out by Nikhef under contract, or under a service level agreement with other organisations, or data arising from commercial or third-party use of Nikhef facilities and installations that are not also part of *Research Activities*. Policy regarding such data is the responsibility of the contracting organisation.

Terms used from RFC 2119

MUST

- This word, or the terms "REQUIRED" or "SHALL", mean that the definition is an absolute requirement of the specification.

SHOULD

- This word, or the adjective "RECOMMENDED", mean that there may exist valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.

RFC219 (BCP 14) - https://www.ietf.org/rfc/rfc2119.txt

Scope

- defines *Research Activities* and re-usable data
- non-reusable data (e.g. collected when building tools/components) need not be subject to the policy
- but most data usually is

Scope

This Policy applies to all research data that are relevant for re-use and produced as a result of Nikhef *Research Activities*, i.e.,

- all approved granted research programmes and granted research projects, and
- research projects so designated and approved by the Nikhef director, and
- any activity that results in *Published* data as per the General Principles.

but there are exclusions

- specific agreements & treaties take precedence (e.g. CERN)

> This Policy shall apply without prejudice to provisions set forth in more specific agreements between Nikhef and any third party, which in all cases take precedence.

- we exclude contract work –
  the responsibility in that case is with the third party
- software as a product in itself *(like FORM, or control software)*
- physical detectors & components *(but archive the design drawing)*
- personal “GDPR” data – *you really ought to think why you need it!*
- administrativa

you *can* obviously still be following the policy, but it will not be as absolutely mandatory

RDM Policy Principles

- Encompasses all kinds of *Data*: raw, derived, published, logs & settings
- Implement NWO Protocol & NWO-I DM Framework
- Be legal, yet use of released data not our concern *(beware of law & NWO)*

- Each *Activity* should have a Data Management Plan "DMP"
  (*already required for NWO, H2020, ERC*)
  – *or* at least implement the 'default' guidance in the Policy
- At least outline of DMP in proposal phase - *already now for NWO/H2020*
- DMP should follow current best practice '***for our domain***'

Open Data Principle

- Data should become public in 6 months following publication
  - unless otherwise agreed (e.g. for the LHC experiments)
  - or if access should be limited to ensure patents, or first publication, or protection of individuals, ethical things, or other good reasons –these reasons to be stated in the DMP
- Data should be *replicatable* and there must be a ‘replication package’

> 10. The *Data* should be made available in a format and manner that allows appropriately qualified and trained researchers in the research domain to replicate the published results.

Data needs Metadata

- Comes out of the FAIR principles
- Is strongly emphasized in the NWO protocol and policy framework

> 12. To enable *Data* to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other qualified researchers to understand the research and re-use potential of the data. *Published* results should include information on how to access the supporting data.

- The '*qualified*' researcher is there for a reason – we do not anticipate citizen science here, but target domain-trained physicists

… or the meta-data should be infinitely detailed and un-doable …

The Data Management Plan

Nikhef recommends that a specific *Data Management Plan* is formulated following the guidance provided by NWO[3] and taking appropriately into account the guidelines and best practices described below. Otherwise, these guidelines and best practices represent the baseline for *Data Management* for the *Research Activity*.

- The specific plan implements the policy – or take the default *and take the time to understand what it means for **your** case:*

1. The *Data Management Plan* or *Research Activity* lead must designate a person or persons responsible for providing the replication package(s) that will be deposited in the repository or repositories.

Nikhef DM Plan Guidance

- guidelines for writing the DMP for each activity

or

- the baseline if you do not write an activity-specific DMP

Describe what will be stored

Plans should cover all *Data* expected to be produced as a result of a project or activity, and that are relevant for re-use or reproducibility, from '*Raw*' to '*Published*'.

Plans should specify which data are to be deposited in a repository, where and for how long, with appropriate justification. The good practice criteria assume that this data is accompanied by sufficient metadata to enable re-use.

Plans may reference the general policy(ies) for the chosen repository(ies) and only include further details related to the specific project. It is the responsibility of the person preparing the data management plan to ensure that the repository policy is appropriate. Where *Data* are not to be managed through an established repository, the *Data Management Plan* will need to be more extensive and to provide reassurance on the likely stability and longevity of any repository proposed.

- but keep costs in mind, and don't store re-createable intermediates

Dumping a replication package

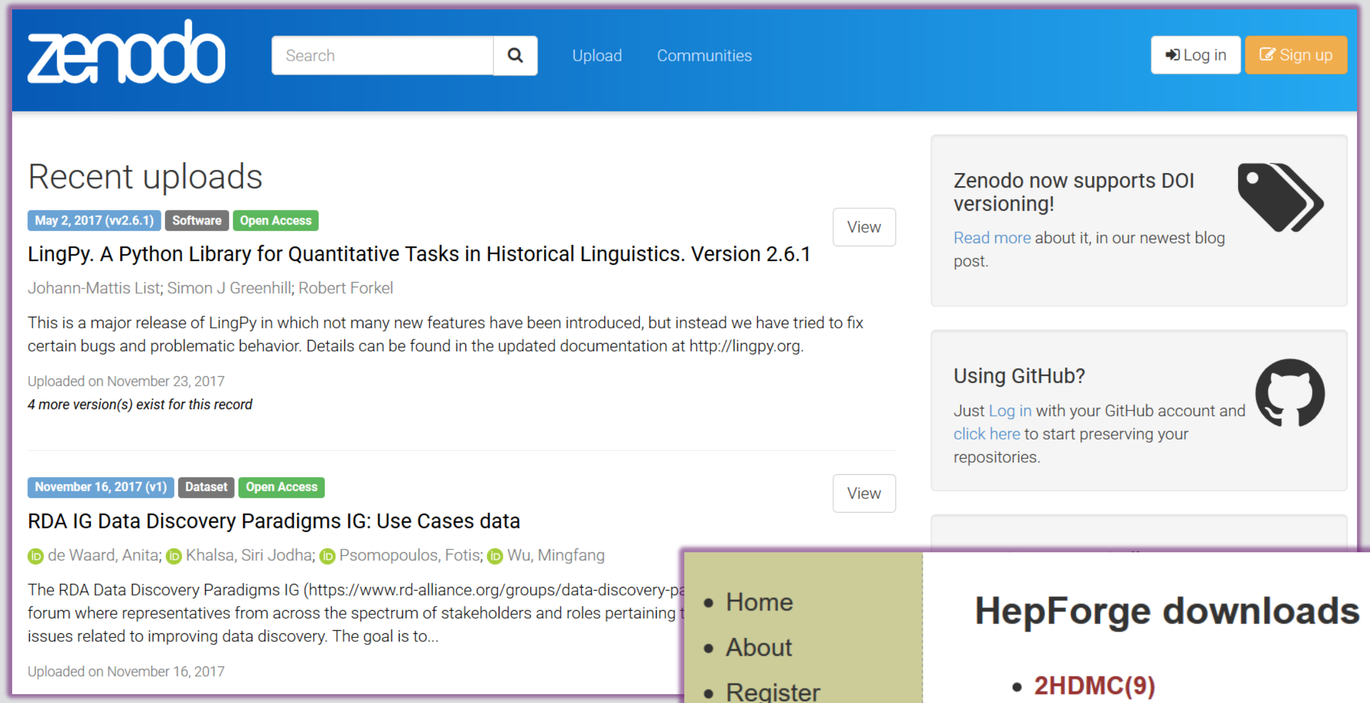
- off-load the curation to an established (non-Nikhef) place

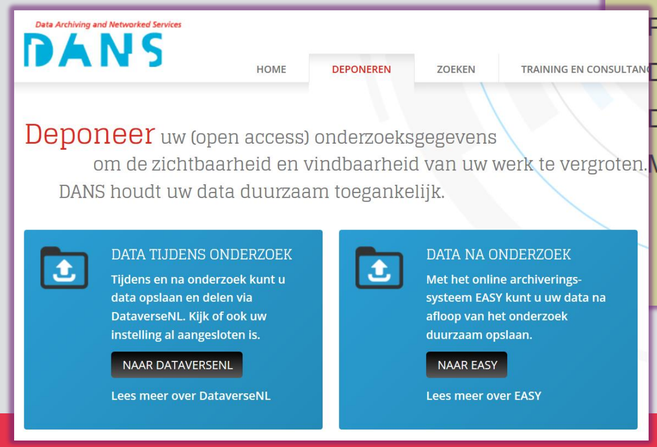
Nikhef would normally expect, upon completion of a research project, programme, or significant phase thereof, the resulting *Data* to be managed through an independently-managed domain-specific repository, a general purpose international repository such as Zenodo[4], a national repository, or a so-designated institutional repository for which long-term sustainability is ensured. The repository(ies) should be chosen so as to maximise the scientific value obtained from aggregation of related data. It may be appropriate to use different repositories for data from different stages of a study.

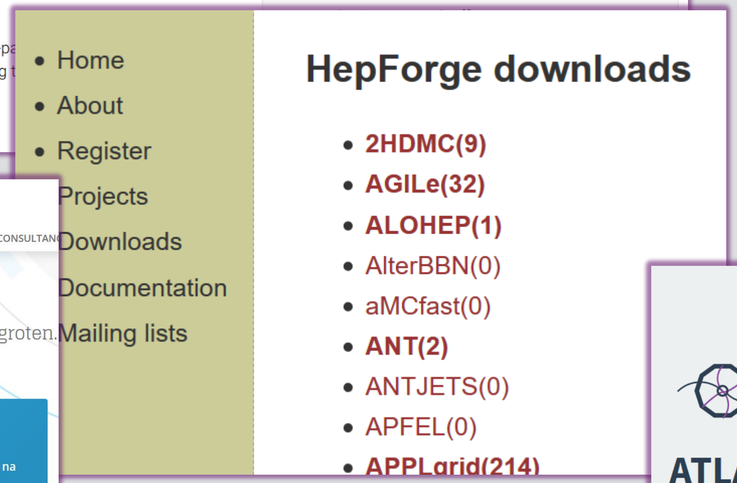
- it's at least a lot cheaper than setting up your own ☺

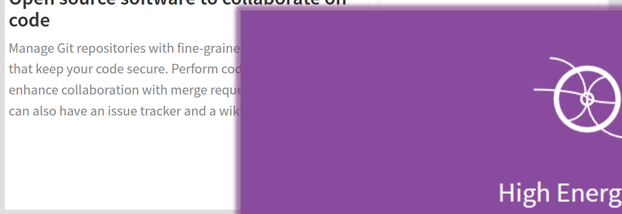
just your own DOI assignment costs €8500/yr for DataCite membership – or €4.50 per DOI @mEDRA

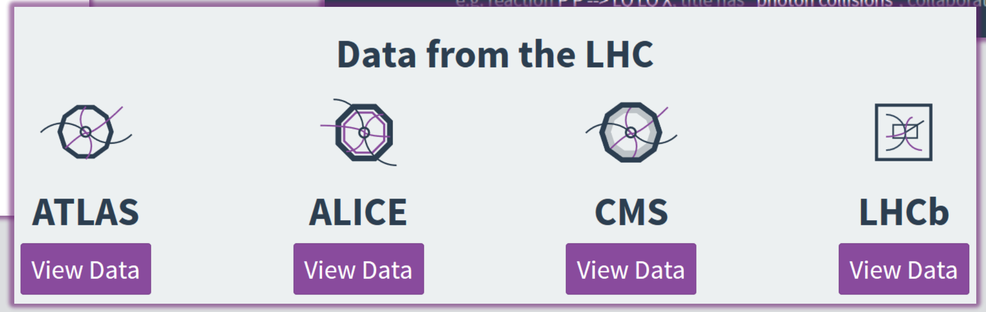
Repositories











Findable – a persistent ID

- Get a DOI (a URL is not usually persistent)
- Link your ORCID to it (bi-directional)

> On deposition of *Data* in a domain-specific, international, or national repository, unique identifiers must be assigned, preferably a DOI, by the researcher or the repository, and the author(s) unambiguously identified with their affiliations, preferably with their ORCID, so that the intellectual contributions of researchers can be acknowledged. Where data is not (yet) fully Open Access, the terms and conditions of access shall be clearly indicated.

- Nikhef cannot assign DOIs itself, we do give URNs, OIDs, and ISBN
- The repositories will do this for you (Zenodo, DANS, FigShare)

just your own DOI assignment costs €8500/yr for DataCite membership – or €4.50 per DOI @mEDRA

Packaging data

- Related data stick together

The deposited *Data* or replication package(s) shall be structured so that *Data* and associated metadata are self-contained and can be referenced by a collective identifier. A *Research Activity* can result in multiple *Data* structures or replication packages, which may be deposited in distinct repositories, based on the nature of the *Data* or the way of publication of the results derived therefrom.

Without prejudice to the requirements of any repository, the *Data* structure or replication package need not be a single object or archive, but may be contained in a structured hierarchy of objects and permanently resolvable references[5]. An appropriate format or structure, such as but not limited to HEPData[6], should be chosen to ensure appropriate metadata, including the Dublin Core Metadata Element Set[7], is registered.

Archive or file hierarchy?

- Submission to a repository usually is in a single file
- But in the Nikhef repository, it might be more useful as a directory
- as long as it contains the meta-data and description

[5] While an format resulting in a single 'file' (e.g. a *tar* archive of Root files, associated scripts, references to on-line notebooks, and a 'readme' file) may be preferred or required for deposition in a public repository, especially during execution of the research or for ease of re-use such a set of Root files, scripts, git urls, &c, could better be maintained in a hierarchical directory structure on Nikhef-provided and bit-curated storage resources. The URL to such a hierarchy on persistent and curated storage is considered a valid unique identifier for internal use if such a path is non-reassigned (e.g. is dated or programmatically generated)

- this should be on backed-up, CT managed, **so-designated** storage!

Meta-data format

Remember the 'rich' meta-data and machine-based research?

Ensure at least minimally useful meta-data is there!

- https://hepdata.net/submission
- https://guidelines.openaire.eu/en/latest/data/index.html
- http://dublincore.org/documents/dces/

Not *that* complicated a set

- **Identifier**
- **Creator**
- **Title**
- **Publisher**
- **PublicationYear**
- Subject (key words)
- *Contributor*
- **Date**
- Language
- ResourceType
- *AlternateIdentifier*
- **RelatedIdentifier**
- *Size*
- *Format*
- *Version*
- **Rights**
- **Description**
- *GeoLocation*

Pick the right path!

- Not all storage is created equal – ask the CT for the QoS

> During execution of the research, *Data* must be managed through and maintained on resources that ensure durability, persistency, and continuity or access. This would normally mean those resources provided centrally by, or by way of, the Nikhef institutional information technology services, or on resources so designated by the research collaboration. The choice of storage quality of service[8] must be commensurate with the value of the data.

- and scan paper log books and notes

> *Data* that are not 'born digitally', such as (written) '*Log*' data, must be put in digital form without undue delay and curated alongside any digital research data.

Immutable results: integrity

- This is a basic property of the required *replication package*

Plans should provide suitable quality assurance concerning the extent to which *Data* can be or have been modified. Where specific data sets are not to be retained, the processes for obtaining such data sets should be specified and conform to the standard accepted procedures within the scientific field at that time.

Wherever possible Nikhef would expect the original data (i.e. from which other related data can in principle be derived) to be retained for the longest possible period, with ten years after the end of the project being a reasonable minimum. For data that by their nature cannot be re-measured, effort should be made to retain them 'in perpetuity'.

Don’t overload the MT

Designs, drawings, and models of experimental apparatus used in the research project are considered sufficient substitute for any physical apparatus. Such designs, drawings, and models in digital form may be preserved in an institutional repository only, and in a representation and format appropriate at the time, even if such a format is proprietary. If a description of the apparatus is published in a manner that permits re-creation by a qualified engineer, the original designs, drawings, and models need not be published.

“managing NX TeamCenter and COMSOL is already complex and both retain traceability”

In Practice

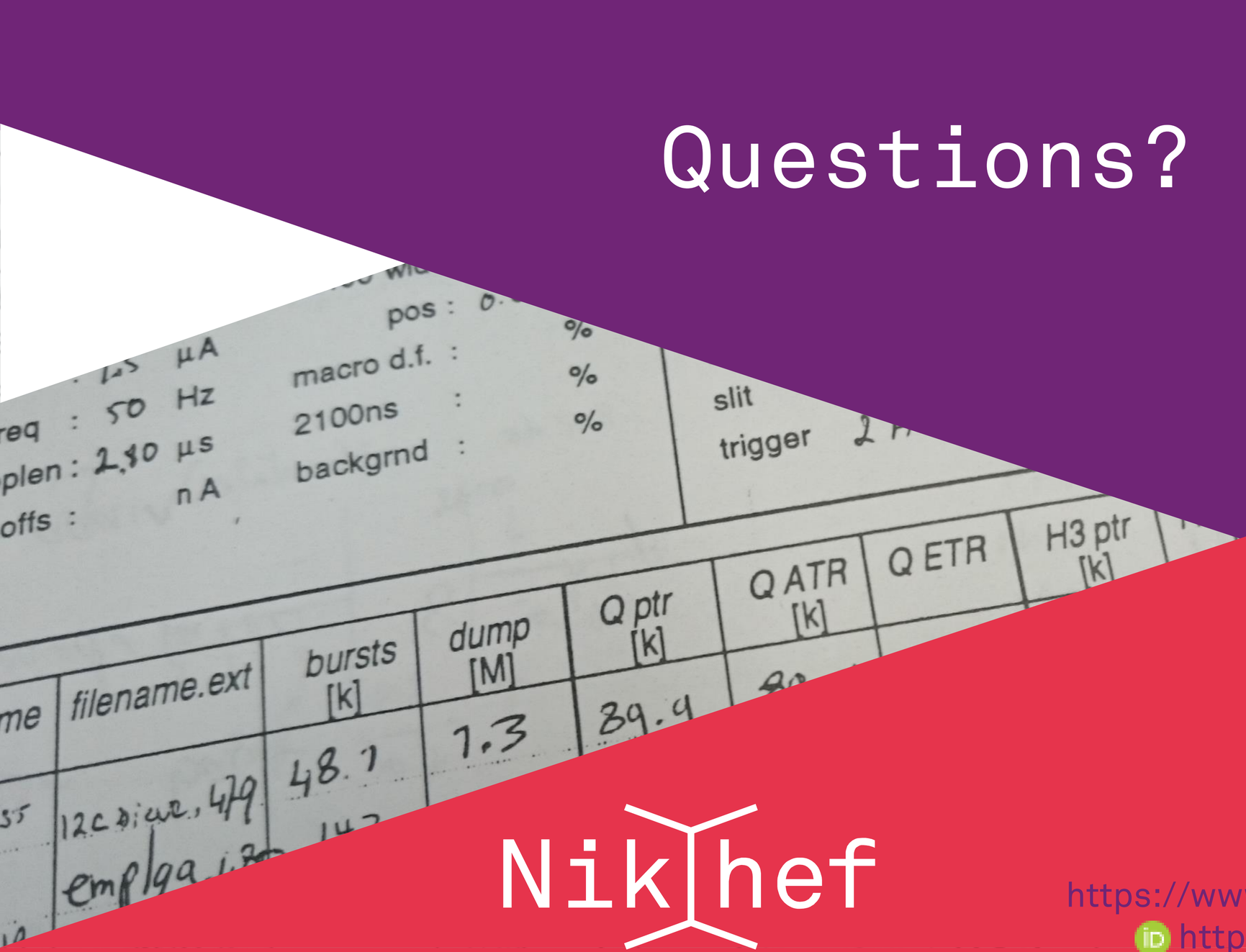
- Every *research activity* will need a responsible person
- Think about what needs to be in the replication package
- Decide if the results can be public – then select the proper repository
- Zenodo takes 50 GBytes/dataset, but unlimited # of datasets ☺
- HEPData takes and assesses data submissions – conforming to 'FAIR'
- For 'local' and proprietary data – CT should be making available space for this (with a QoS like /project) – *don't modify content there*
- We do need a registry for research output in our local repo (and need that anyway for the Annual Report and reviews)
- Indicate Nikhef as *publisher*, and you as *creator* – and include your ORCID

In Practice

- Review the Computing Course training on research data
  Jeff gave one on Nov 27th
- Use DMP copy-pasting from *https://www.nikhef.nl/grid/nikhef/dmp/*
- Ask the CT/PDP team for help [rdm-support@nikhef.nl]
  - if you have to cook up a DMP for NWO
  - if you want to re-use a stock DMP for the LHC experiments
  - if you need some GDPR guidance (but: IANAL!)
- This policy and practice *needs to evolve* and *will change*, but for now

Sit back, relax, and enjoy your data!

Questions?



Nikhef

David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606