Nikhef RDM Policy – first experiences

Contactgroep DM NWO-I

Nikhef

David Groep
davidg@nikhef.nl

Nikhef RDM Development Process

A long journey – and for long at a sufficiently abstract level

- prominently present in H2020/ERC (and in NWO now)
- participated in the FOM DM pilot for *Projectruimte* proposals (2016)
- schemes were adaptable to domain-specific needs

August 2016

"*NWO Institute Data Management Policy Framework*" became binding

- development of institute policy now must fit within this framework
- flexibility needs to be found within those constraints (keeping cost in mind)

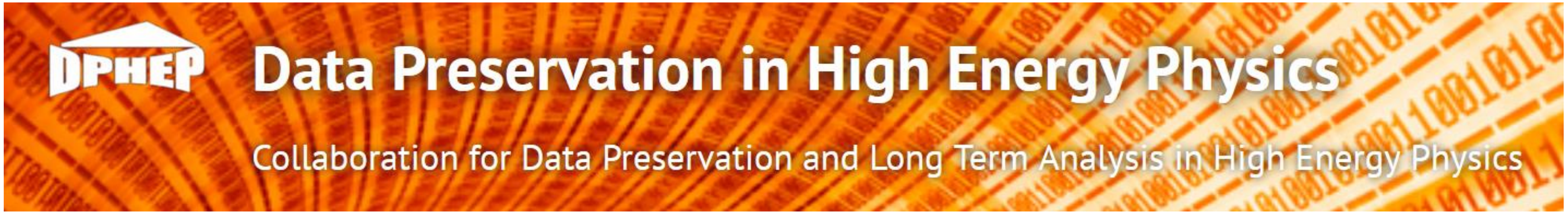
Summer 2017

- decide on a hierarchical approach, separating policy and practice statements
- with a sensible default for 'small' activities (but that is still a challenge)

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

http://dphep.org/

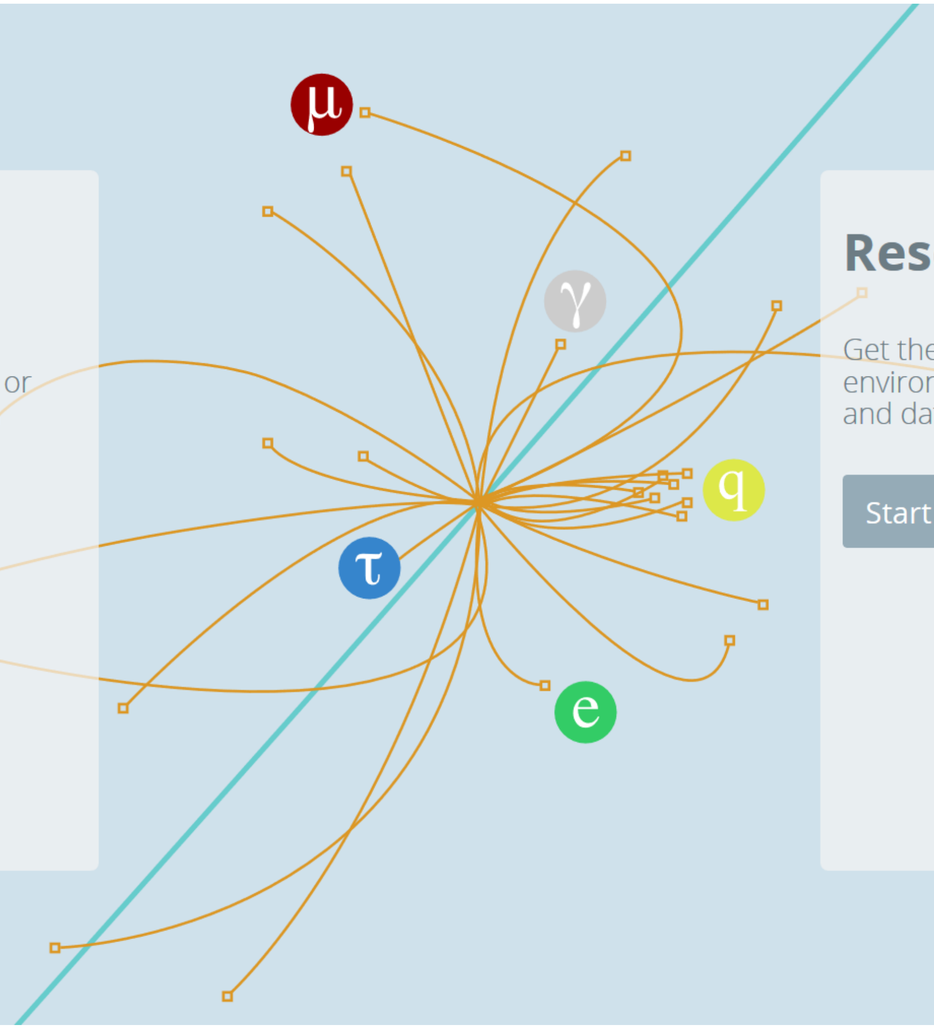
opendata.cern.ch

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

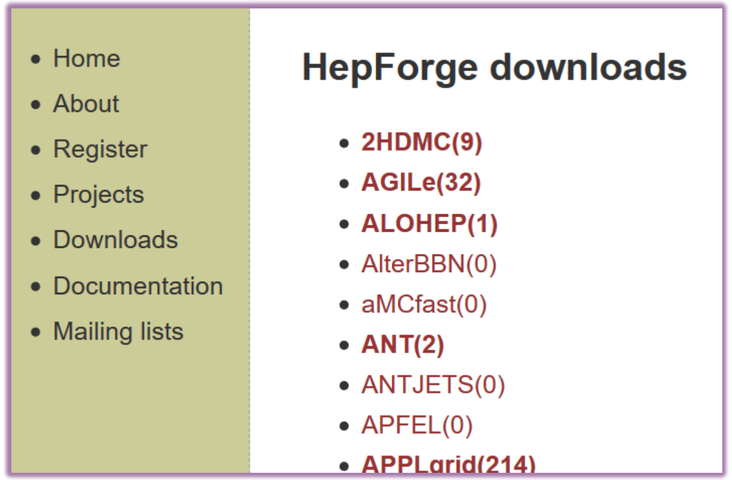
According to the ATLAS Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

LHCb

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

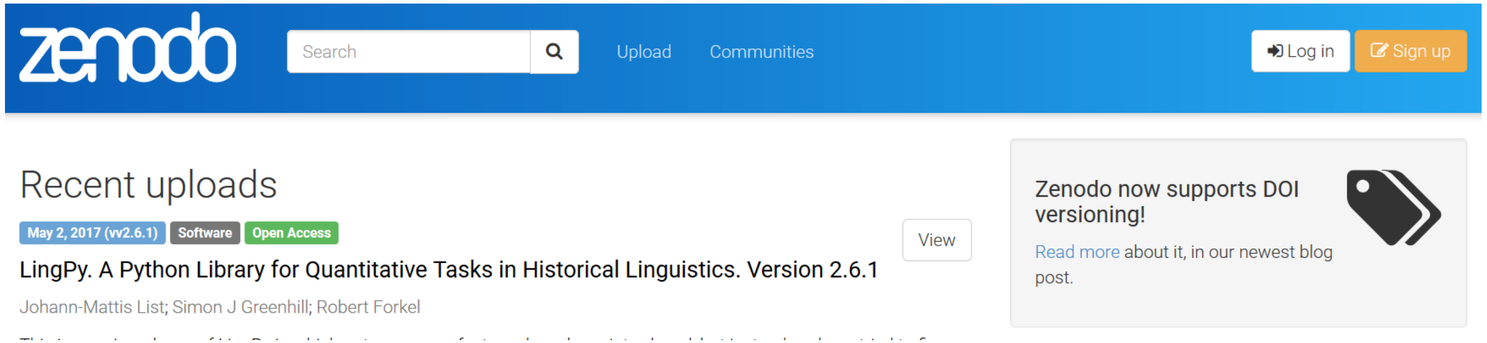
We have a diverse tradition …

- DPHEP
- HEPForge (and many sub-repositories)
- HepData
- CERN OC3 archive
- INSPIRE-HEP ('SPIRES')



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

- Zenodo



https://www.nikhef.nl/grid/nikhef/dmp/

Nikhef RDMP ToR

- must implement the NWO-I DM Framework (and NWO Protocol)
  - *leaves us freedom of implementation, but we signed it, and it does require specific elements like the Replication Package*
- not undermine any past or future data management customs already in use
- 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

Re-use as much as possible from partners – we re-used STFC/RAL ☺

The Document ...

Nikhef Research Data Management Policy v02

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.

Uses RFC2119 terminology

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.

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.

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

- Supply outline of DMP in proposal phase - already now for NWO/H2020
- DMP SHOULD follow current best practice '**for our domain**'

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 control software)*
- physical detectors & components *(but archive the design drawing)*
- personal “GDPR” data – *you really ought to think why you need it @Nikhef*
- administrativa

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

Standard DMPs

- Many global experiments have a DM policy already in place: LHC experiments, Auger, LVC, ...
- To answer the NWO DM paragraph, we developed standard templates

https://www.nikhef.nl/grid/nikhef/dmp/

Form Data Management Plan

1.	**Administrative information**			
1.1	Project number	ENTER YOUR PROJECT NO	ENTER YOUR PROJECT NO	ENTER YOUR PROJECT NO
2.	**Description data set**			
2.1	Describe the data that will be collected/generated	Based on the definition of "relevant data" above, we can take the "ATLAS Policy on Data Preservation" [https://cds.cern.ch/record/2012333] to define the "wide consensus"; this policy is established within the context of the DPHEP study group of ICFA (see answer 3.6 for details). The project is based on data collected by the Atlas (www.atlas.ch) experiment at CERN (www.cern.ch). All relevant data generated by the project will become part of the "ATLAS Data" as defined in ATLAS Policy on	Based on the definition of "relevant data" above, we can take the "ALICE data preservation stategy" [http://opendata.cern.ch/record/412] to define the "wide consensus"; this policy is established within the context of the DPHEP study group of ICFA (see answer 3.6 for details). The project is based on data collected by the ALICE (aliceinfo.cern.ch) experiment at CERN (www.cern.ch). All relevant data generated by the project will become part of the ALICE data (i.e. the ensemble of data in all data formats listed as ALICE data formats) as defined in the ALICE data preservation strategy as endorsed by its Collaboration Board on 6th of	Based on the definition of "relevant data" above, we can take the "LHCb external data access policy" [http://opendata.cern.ch/record/410] to define the "wide consensus"; this policy is established within the context of the DPHEP study group of ICFA (see answer 3.6 for details). The project is based on data collected by the LHCb (lhcb.cern.ch) experiment at CERN (www.cern.ch). All relevant data generated by the project will become part of the LHCb data (i.e. the ensemble of data in all data formats listed as LHCb data formats) as defined in the LHCb data

Otherwise, leverage the community

- The aim for all non-standard DMP is to leverage existing community services and initiatives (DPHEP and more)

> 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.

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

Repositories

zenodo

Search | Upload | Communities | Log in | Sign up

Recent uploads

May 2, 2017 (v2.6.1) | Software | Open Access | View

LingPy. A Python Library for Quantitative Tasks in Historical Linguistics. Version 2.6.1

Johann-Mattis List; Simon J Greenhill; Robert Forkel

This is a major release of LingPy in which not many new features have been introduced, but instead we have tried to fix certain bugs and problematic behavior. Details can be found in the updated documentation at http://lingpy.org.

Uploaded on November 23, 2017

4 more version(s) exist for this record

November 16, 2017 (v1) | Dataset | Open Access | View

RDA IG Data Discovery Paradigms IG: Use Cases data

de Waard, Anita; Khalsa, Siri Jodha; Psomopoulos, Fotis; Wu, Mingfang

The RDA Data Discovery Paradigms IG (https://www.rd-alliance.org/groups/data-discovery-pa
forum where representatives from across the spectrum of stakeholders and roles pertaining t
issues related to improving data discovery. The goal is to...

Uploaded on November 16, 2017

Zenodo now supports DOI versioning!

Read more about it, in our newest blog post.

Using GitHub?

Just Log in with your GitHub account and click here to start preserving your repositories.

GitLab Community Edition

Open source software to collaborate on code

Manage Git repositories with fine-grained access controls that keep your code secure. Perf
enhance collaboration with mer
can also have an issue tracker a

Sign in | Register

Username or email

Password

- Home
- About
- Register
- Projects
- Downloads
- Documentation
- Mailing lists

HepForge downloads

- 2HDMC(9)
- AGILe(32)
- ALOHEP(1)
- AlterBBN(0)
- aMCfast(0)
- ANT(2)
- ANTJETS(0)
- APFEL(0)
- APPLgrid(214)

HEPData

High Energy Physics Data Repository

This new site replaces the old site at **http://hepdata.cedar.ac.uk**.

Search on 8566 publications and 71328 data tables.

Search for a paper, author, experiment, reaction | Search | Advanced

e.g. reaction P P --> L0 L0 X, title has "photon collisions", collaboration is LHCf or D0.

Data from the LHC

ATLAS	ALICE	CMS	LHCb
View Data	View Data	View Data	View Data

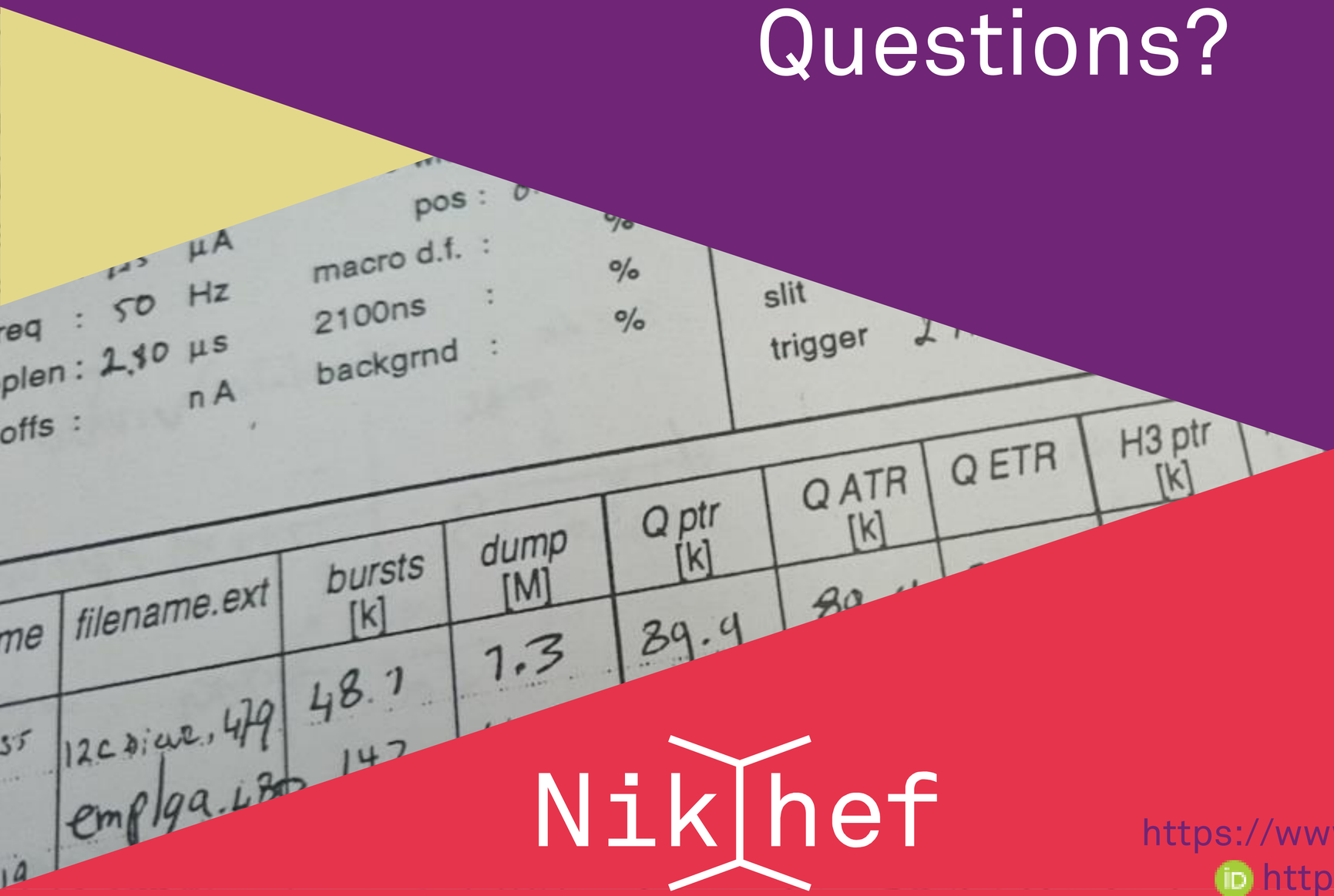
B2SAFE

B2HANDLE

Early feed-back from the staff

- **Standard DMP templates are much appreciated**
  - Need to take balanced care about RDM in MoU negotiations
    'seek to ensure' – not require
- **Wider diversity in practices for 'smaller' collaborations**
  - many concerns about feasibility and cost/benefit analysis
  - 'depth' of the replication package is a big concern
  - the "I" *premise* of FAIR does not quite work in our domain
- **Policy implementation will be a phased approach**
- **Training is an essential component**
  - both on RI and log keeping (e.g. Jupyter notebooks)
  - and on proper use of existing repository facilities (storage QoS classes)

Questions?



Nikhef

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

The “FAIR” Principles

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