Nikhef

Maastricht University

Identifiers, repositories, licenses, and DMPs

When your data becomes part of something bigger

David Groep
Nikhef
Computing Course 2023

Objectives for this 'data and software' session

- Know common persistent identifier schemes; such as DOI, hndl.net and ORCID

- Possess an ORCID identifier and
  know how to update basic information in their ORCID record

- Know the importance of applying licenses to data and software for re-usability

- Know the basic licenses for data and software, and
  know where to find guidance on license application,
  and the main differences between licenses

- Know the basic outline of a data management plan

- Know how to fill in a mock data management plan using the on-line tools
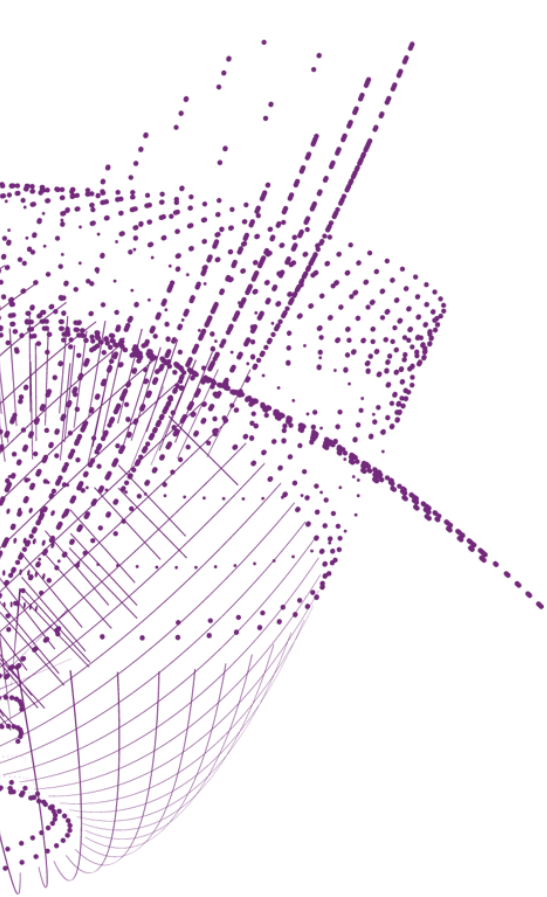
Persistent identifiers
for publications, data … and for researchers

I'm not a number!

When your data becomes part of something bigger

Getting a unique identifier is hard … which Anna Wilson?

First Name	Last Name	Other Names	Affiliations
Anna	Wilson		University of Alberta, Weir Memorial Law Library University of Alberta
Anna	Wilson		Phi Beta Kappa Society, Washington University in St Louis, Washington University in St Louis School of Medicine
Anna	Wilson		Charles Sturt University, Premier Specialists, St George Hospital, UNSW Sydney, University of Wollongong
Anna	Wilson		Duke University, University of California San Diego Scripps Institution of Oceanography
Anna	Wilson		Lund University, Lund University Samhällsvetenskapliga fakulteten, University of Lausanne, Universität St. Gallen
Anna	Wilson		Dartmouth College, U.S. Geological Survey, University of New Hampshire
Anna	Wilson		Hennepin Healthcare Research Institute, University of Minnesota
Anna	Wilson	A N Wilson, A Wilson, Anna N Wilson	Abertay University, Australian National University, University of Bristol, University of Canberra, University of Glasgow, University of Liverpool, University of Oxford, University of Stirling, University of York, Yale University
Anna	Wilson		Harvard University
Anna	Wilson		Auburn University
Anna	Wilson		University of Salford

When your data be

Assigning a globally unique non-reassigned one helps:

ORCID ID	First Name	Last Name	Other Names	Affiliations
0000-0003-2397-7941	Anna	Wilson		University of Alberta, Weir Memorial Law Library University of Alberta
0000-0001-6285-3824	Anna	Wilson		Phi Beta Kappa Society, Washington University in St Louis, Washington University in St Louis School of Medicine
0000-0001-5596-2109	Anna	Wilson		Charles Sturt University, Premier Specialists, St George Hospital, UNSW Sydney, University of Wollongong
0000-0001-7342-1955	Anna	Wilson		Duke University, University of California San Diego Scripps Institution of Oceanography
0000-0002-4478-675X	Anna	Wilson		Lund University, Lund University Samhällsvetenskapliga fakulteten, University of Lausanne, Universität St. Gallen
0000-0002-9737-2614	Anna	Wilson		Dartmouth College, U.S. Geological Survey, University of New Hampshire
0000-0002-4543-1344	Anna	Wilson		Hennepin Healthcare Research Institute, University of Minnesota
0000-0001-6928-1689	Anna	Wilson	A N Wilson, A Wilson, Anna N Wilson	Abertay University, Australian National University, University of Bristol, University of Canberra, University of Glasgow, University of Liverpool, University of Oxford, University of Stirling, University of York, Yale University
0000-0002-5229-9716	Anna	Wilson		Harvard University
0000-0002-8575-7138	Anna	Wilson		Auburn University
0000-0002-5563-2318	Anna	Wilson		University of Salford

What should an identifier scheme do?

- **unique**
- **persistent**
- **non-reassigned**

- findable: identifier should be good enough to take you to the object

- for 'evolving' objects: be able to identify a *collection* (and 'latest version')

- come from an *authoritative source*

When your data becomes part of something bigger

Not all identifiers are created equal

Technical qualities, but also 'impact perception'

- ObjectIDs (OID)
- Universal Resource Names (URN)
- Digital Object Identifiers (DOI)
- Handles (hndl.net)

And then there are plenty of *non-persistent identifiers*, like URLs

When your data becomes part of something bigger

Uniform Resource Indicators == URLs + URNs

But URLs *do* change for no good reason (or simply because "functions follows form")

W3C *Style*

Cool URIs don't change

What makes a cool URI?
A cool URI is one which does not change.
What sorts of URI change?
URIs don't change: people change them.

There are no reasons at all in theory for people to change URIs

URNs (and OIDs) are unique + persistent but hard to resolve – just try find the path to
`urn:geant:nikhef.nl:idm:md:entity:spproxy:201606`

Tim Berners-Lee, https://www.w3.org/Provider/Style/URI, 1998; for the URN namespace, see RFC 4926, then http://www.dante.net/urn-geant/urn-geant.html, then see #12

When your data becomes part of something bigger

DOI and Hndl.net

DOIs – a persistent link (opaque) to publications & more …

- originally come from the publishing industry (CrossRef)
- perception is still very much 'high quality paper' like
- libraries and evaluation reports really love DOIs
- but are a bit expensive for repositories with lots of objects

but there are now many large-scale repositories for data, whitepapers, drafts, presentations, &c that assign DOIs

- Zenodo (hosted at CERN and supported by OpenAIRE)
- 4TU.RD (hosted at TU Delft)
- and commercial services like Figshare

and now the movie industry even assigns DOIs to films and broadcasts

One well-cited dataset

https://doi.org/10.7935/K5MW2F23

but an updated version (in this case a phase correction) results in a **new** DOI
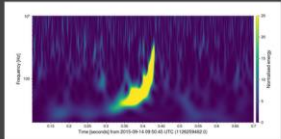
https://doi.org/10.7935/82H3-HH23

This is the data set corresponding to GW150914
data reference https://www.gw-openscience.org/eventapi/html/GWTC-1-confident/GW150914/v3 - also adhered to common data formats (here: HDF5) – see later!

DOI:10.7483/OPENDATA.ATLAS.CPVE.5FA9

MC:Zττ + Jets, for 2016 ATLAS open data release

Zττ + Jets, ATLAS Collaboration

Cite as: ATLAS Collaboration (2016). MC:Zττ + Jets, for 2016 ATLAS open data release. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.CPVE.5FA9

`Dataset` `Simulated` `ATLAS` `8TeV` `CERN-LHC`

mc_147772.Ztautau.root   91.4 MB   ⬇ Download

Disclaimer

The open data are released under the Creative Commons CC0 waiver. Neither ATLAS nor CERN endorse any works, scientific or otherwise, produced using these data. All releases will have a unique DOI that you are requested to cite in any applications or publications.

![PUBLIC DOMAIN]

http://opendata.cern.ch/record/3822, resolved from the DOI http://doi.org/10.7483/OPENDATA.ATLAS.CPVE.5FA9

When your data becomes part of something bigger

DOIs can be assigned to almost any object

https://zenodo.org/ and for your own experimentation, use https://sandbox.zenodo.org/

When your data becomes part of something bigger

Make sure you have identifiers for everything

Also for data sets, software, standards specifications, &c
– they come in handy for grant applications, like for a Veni:
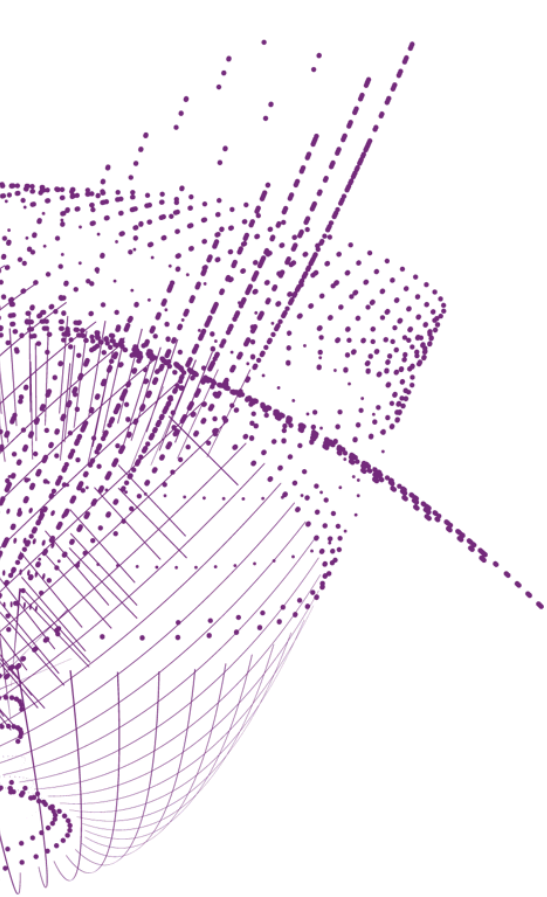
4b. Key output

Provide the references to your key output (max. 10) and add a motivation for the selection of each of these items. Please number the items consecutively. You are allowed to use one hyperlink per item, which refers directly to the output (**e.g. a DOI**). You may not mention H-indexes, journal impact factors, or any other indicator or term that refers to the general quality or reputation of a journal, publisher, or publication platform, rather than to the individual output item. For more information, expand the Explanatory Notes.

11. arXiv:nucl-ex/0703007 [pdf, ps, other] nucl-ex doi 10.1016/j.physletb.2007.08.034
16O(e,e'p) reaction at large missing energy
Authors: M. Iodice, E. Cisbani, R. De Leo, S. Frullani, F. Garibaldi, D. L. Groep, W. H. A. Hesselink, E. Jans, J. G. Onderwater, R. Perrino, J. Ryckebusch, R. Starink, G. M. Urciuoli
Abstract: We investigate the origin of the strength at large missing energies in electron-induced proton

- for papers, journal will assign one

- arXiv links are persistent as well and link to later DOIs

From the Veni 2022 grant template, see https://www.nwo.nl/en/calls/nwo-talent-programme-veni-science-domein-2022

When your data becomes part of something bigger

Identifying you …

When your data becomes part of something bigger

ORCID – the Open Reseacher and Contributor ID

- helps to uniquely identify your papers
- makes sure recognition goes to the right author
- helps you build your list of publications

- can act as an academic resume
- use it to login to R&E services

- needed for grant applications

Go to https://orcid.org/
Example from NWO Roadmap 2021 Application Form, from https://www.nwo.nl/en/calls/large-scale-research-infrastructure-lsri-national-roadmap-consortia-2021

When your data becomes part of something bigger

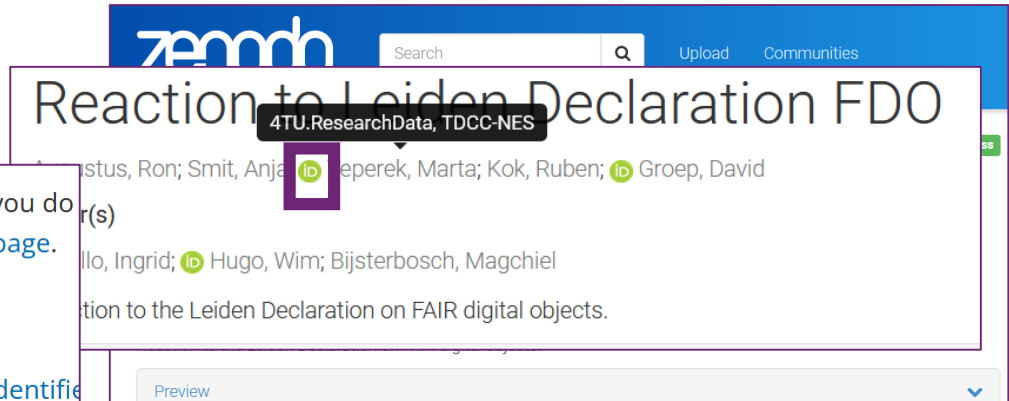# Recognising and using ORCID with linking

- based on the ISNI format

orcid.org/**0000-0003-1026-6606**

- can be linked automatically in journals and repositories

We encourage all arXiv authors to link their ORCID iD with arXiv. If you do process. Once completed you will see your ORCID iD on your user page.

iD **Link my arXiv account with ORCID**

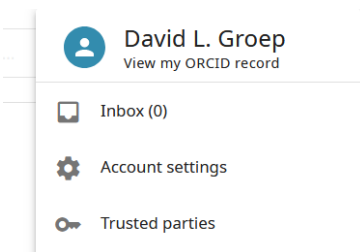arXiv will use ORCID iDs in preference to the internal arXiv author identifie

zenodo   Search   Upload   Communities

Reaction to Leiden Declaration FDO

4TU.ResearchData, TDCC-NES

...stus, Ron; Smit, Anja...eperek, Marta; Kok, Ruben; iD Groep, David

...lo, Ingrid; iD Hugo, Wim; Bijsterbosch, Magchiel

...tion to the Leiden Declaration on FAIR digital objects.

Preview

https://arxiv.org/help/orcid; https://doi.org/10.5281/zenodo.7260200

When your data becomes part of something bigger

Authenticating to your ORCID

Your ORCID ID is *for life* – not only when you're at Nikhef
- you can link multiple *authentication sources* to your ORCID
- one is a username-password specific to the service
- you can link one (or more) institutional IDs *(also most universities should work)*

David L. Groep
View my ORCID record

Inbox (0)

Account settings

Trusted parties

Alternate sign in accounts

You can sign into ORCID using the personal and institutional accounts you have linked to your ORCID record.

Learn more about using alternate accounts to sign in to ORCID

Account	Alternate sign in ID	Access granted	
IGTF Certificate Proxy	David Groep davidg@nikhef.nl	2017-03-08	🗑
Nikhef	davidg@nikhef.nl	2016-05-09	🗑

- you should add *multiple email addresses* to your ORCID (also a personal one) – these are *not public by default!*
- as well as multiple login methods! At some point, you may leave your home org!

Multiple ways into your ORCID is good

Access through your institution

You may sign into the ORCID Registry using institutional accounts you already have, like one from your university. If you don't already have an ORCID iD, you will be prompted to create one. **Learn more about different ways to sign in to ORCID.**

institution logo

Organization's name

CERN   ✕

Go back   **CONTINUE**

Link your CERN account to your ORCID record

You are signed into **CERN** as **David Groep**

To finish linking this CERN account to ORCID, sign into your ORCID iD below. You will only need to complete this step once. After your account is linked, you will be able to access your ORCID record with your CERN account. Questions? **Visit our knowledge base**

Email or 16-digit ORCID iD

0000-0003-1026-6606

example@email.com or 0000-0001-2345-6789

Password

••••••••••••••••••••••••

Sign in and link your CERN account

Cancel and go back

Forgot your password or ORCID ID?

Don't have an ORCID iD yet? **Register now**

https://orcid.org/account

∨ Alternate sign in accounts

You can sign into ORCID using the personal and institutional accounts you have linked to your ORCID record.

Learn more about using alternate accounts to sign in to ORCID

Account	Alternate sign in ID	Access granted	
IGTF Certificate Proxy	David Groep davidg@nikhef.nl	2017-03-08	🗑
CERN	groep@cern.ch	2022-11-12	🗑
Nikhef	davidg@nikhef.nl	2016-05-09	🗑
Maastricht University	P70081609@unimaas.nl	2023-03-08	🗑

When your data becomes part of something bigger

Linking your publications to ORCID
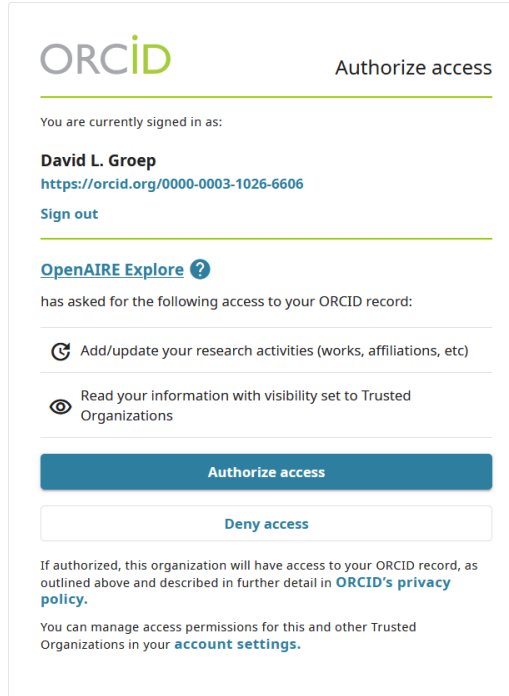
Import from Scopus and commercial publishers

logging in to Zenodo with your ORCID

Some agencies can *update* your record for you

You do get a notice, like this:

When your data becomes part of something bigger

Now go over to orcid.org, and …

When your data becomes part of something bigger

And sign up for Zenodo with ORCID or github (or password)

When your data becomes part of something bigger

And on Djehuty – the Nikhef RDM Repository

https://archive.nikhef.nl

When your data becomes part of something bigger

But what about PID and software?
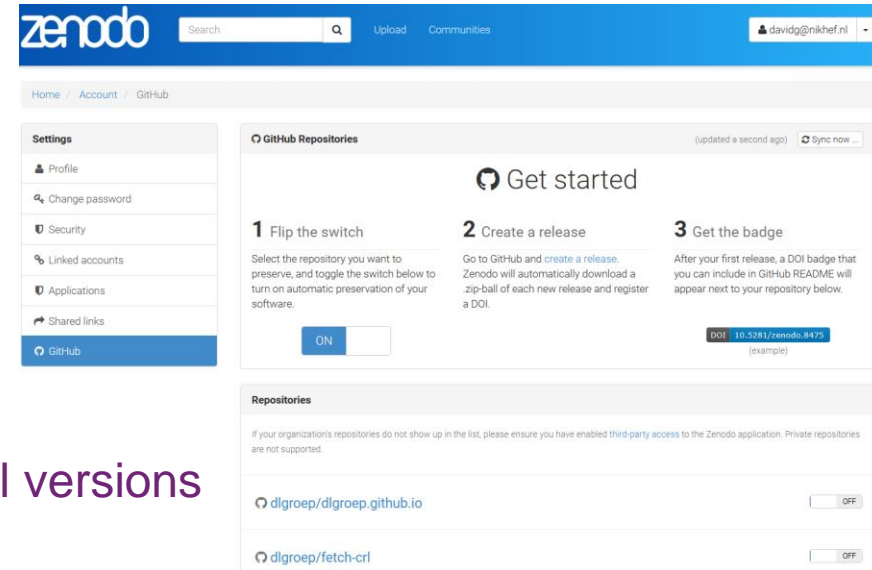
A personal github repository is not a persistent identifier either

- it is good for FOSS and collaboration, but it is not a publication … *yet*

- from Zenodo, link your github account
- in Zedono select your repository
- in Github, make a *release*

this will be automatically uploaded into Zenodo – and you have persistent DOIs for all versions

specifically useful for research software and your reproduction package

https://zenodo.org/account/settings/github/ and https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content
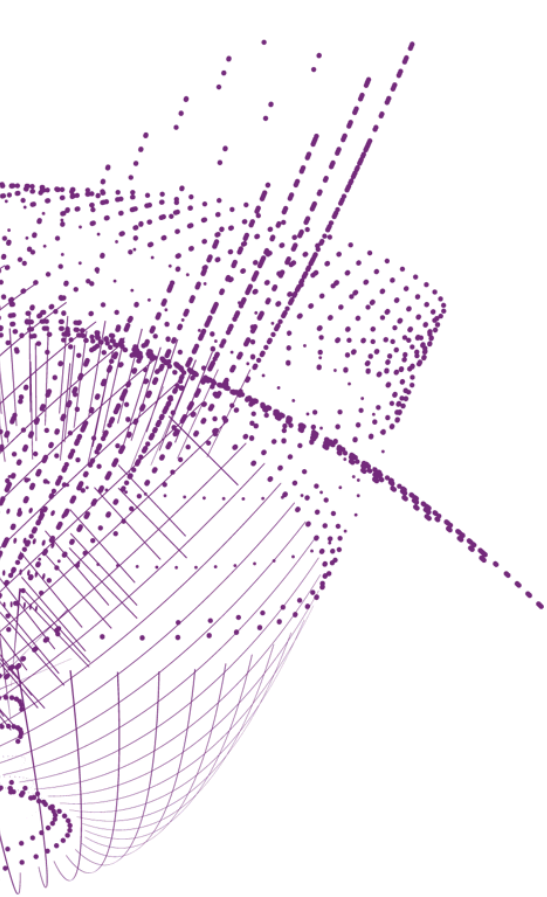
Licensing and open source

Enabling Software Re-use

When your data becomes part of something bigger
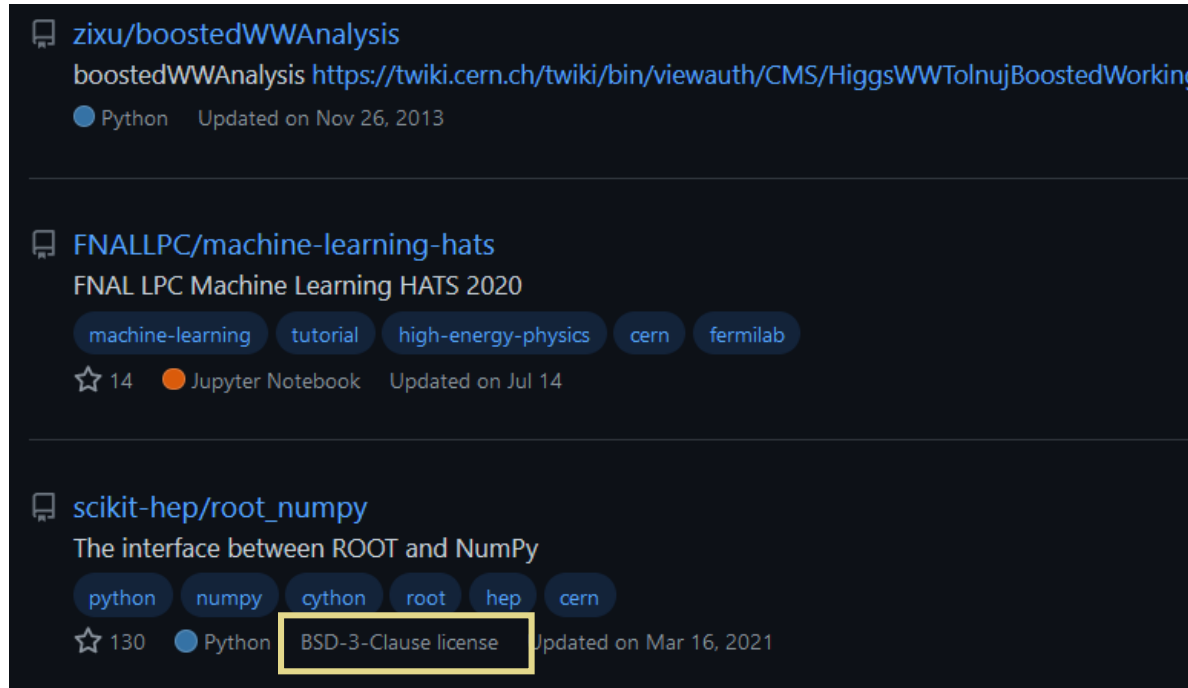
however: code and source files without a license means 'all rights reserved' ☹

When your data becomes part of something bigger

What do we actually want to achieve with our software?

meet open science needs

re-use & replication

collaborative development

not encourage 'patent hostages'

make generally available the results of its experimental and theoretical work*

be understandable

When your data becomes part of something bigger

E.g. Article II.1 CERN convention, see https://council.web.cern.ch/en/content/convention-establishment-european-organization-nuclear-research

A range of Open Source licenses to choose from

BSD family
compact licenses
(BSD 3-clause,
BSD 2-clause, MIT)

re-use friendly | no patent protection

contributions unspecified

pre-Apache family
(GEANT4, EDG)

re-use friendly | no patent protection

contributions auto-imported

Mozilla family
(MPL, Perl Artistic)

re-use friendly | patent protection

contributions unspecified

Apache family
(Apache 2.0)

re-use friendly | patent protection

contributions auto-imported

copyleft family
(GPL)

re-use unfriendly | 3.0+patent protection

contributions auto-imported: N/A

lesser copyleft
(LGPL)

re-use unfriendly | no patent protection

contributions auto-imported: N/A

https://opensource.org/licenses

When your data becomes part of something bigger

Nikhef

Licenses cannot be mixed at random …

For example, Apache 2.0 and GPL 2.0 are 'incompatible' (because Apache protects against software patents, which is a restriction beyond GPLv2), but Apache 2.0 can be linked in GPL **v3** software.

But not the other way round: GPL software is infectious / viral in nature

"Apache 2 software can therefore be included in GPLv3 projects […]. However, GPLv3 software cannot be included in Apache projects. The licenses are incompatible in one direction only, and it is a result of […] the GPLv3 authors' interpretation of copyright law."

https://www.apache.org/licenses/GPL-compatibility.html

When your data becomes part of something bigger

Many LHC experiments have standard license & IP clause

although some LHC experiments are completely silent on this
(*and the CERN Convention, in II.1, does not help in case of IP from contributors*)

> as this single copyright administering entity for the benefit of the LHCb collaboration. This arrangement assists LHCb in achieving the widest possible dissemination and use of its software [3]. The copyright statement that is applied to all LHCb centrally distributed application software is therefore:
>
> "(C) Copyright CERN for the benefit of the LHCb collaboration"

When your data becomes part of something bigger

https://twiki.cern.ch/twiki/pub/LHCb/LHCbNationalComputingBoard/20210215_LHCbSoftwareCopyrightAndLicense.pdf

But LHCb early on got infected with GPL code …

"The LHCb software depends on packages licensed under the "GNU General Public License (GPL)". The terms of GPL require that derivative works be licensed under the same license that governs the original software when distributed.
Accordingly, the LHCb software stack also needs to be licensed under the

"GNU General Public License v3"

[…] the LHCb collaboration – acting through CERN as the copyright holder – can re-license or distribute the software under a dual license scheme, always taking into account dependencies and license compatibility

When your data becomes part of something bigger

Nik|hef

Yet ATLAS is all Apache 2.0, so …

So one common piece of software (Gaudi), which has
no dependencies on GPL or other LHCb core,
is Apache 2.0 licensed
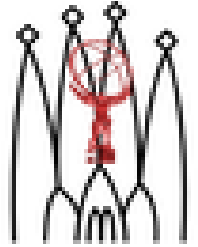
And some independent code, like *Allen*, is also Apache 2.0,
with those bits of general code from LHCb used therein
being *dual licensed*.

https://twiki.cern.ch/twiki/pub/LHCb/LHCbNationalComputingBoard/20210215_LHCbSoftwareCopyrightAndLicense.pdf

Listing contributors

'a successful community has many contributors!'

'… but listing them all will then be a challenge!'

- co-shipped 'contributors' file, or a web page listing contributors
- "members of the XXX collaboration"
  + a web page is commonly used
- some projects list main contributors
  and have just given up

Geant4 Software License

Version 1.0, 28 June 2006

Copyright (c) Copyright Holders of the Geant4 Collaboration, 1994-2006.
See http://cern.ch/geant4/license for details on the copyright holders. All right

- probably worst thing to do is
  to also accept changes
  in the copyright license statement itself
  (the "SymPy" case)

'right to be identified as an author' is a 'moral' right you cannot get rid of, but can be (partially) waived, e.g. as part of employment …

When your data becomes part of something bigger   Source: from the GEANT4 web pages at
https://geant4.web.cern.ch/license/LICENSE.html

… and the long list of contributors

either you get lists like on the left
(and then GEANT4 is a 'small' project)
or you become creative, like
use github's contributor log (e.g. for SimpleSAMLphp)
or link to your project or collaboration page

The most simple open source license: 3-clause BSD

- listing all contributors in the copyright line
- all rights assigned to the organisations (not individual employee)
- but:
  no patent protections

```
29 lines (23 sloc)   1.51 KB

 1   BSD 3-Clause License
 2
 3   Copyright (c) 2009-2017, AARNet, Belnet, HEAnet, SURFnet, UNINETT
 4   All rights reserved.
 5
 6   Redistribution and use in source and binary forms, with or without
 7   modification, are permitted provided that the following conditions are met:
 8
 9   * Redistributions of source code must retain the above copyright notice, this
10     list of conditions and the following disclaimer.
11
12   * Redistributions in binary form must reproduce the above copyright notice,
```

When your data becomes part of something bigger

Have your pick …

Popular Licenses

The following OSI-approved licenses are popular, widely used, or have strong communities:

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license
- Mozilla Public License 2.0
- Common Development and Distribution License
- Eclipse Public License version 2.0

open source
initiative
Approved License®

from: https://opensource.org/licenses

Nikhef

A last word about patents …

… usually for mutual litigation, but have been used against open source (although rejected)

Some licenses try to address that by voiding themselves if the licensee institutes patent litigation involving (parts of) the work against anyone else:

- Apache 2.0, GPL 3.0, Perl Artistic

(11) Application No. **AU 2001100012 A4**
(19) **AUSTRALIAN PATENT OFFICE**

(54) Title
     Circular transportation facilitation device

| (21) | Application No: **2001100012** | (22) | Date of Filing: **2001.05.24** |

(43) Publication Date: **2001.08.02**

(71) Applicant(s)
     John Keogh

(72) Inventor(s)
     Keogh, John Michael

(74) Agent/Attorney
     Sandercock Cowie 69 Robinson Street Dandenong Victoria AU

Australian (light-weight) Innovation Patent #2001100012, from 2001, since voided after international upheaval ☺
http://pericles.ipaustralia.gov.au/ols/auspat/applicationDetails.do
plus the 2001 Ignobel prize, of course!
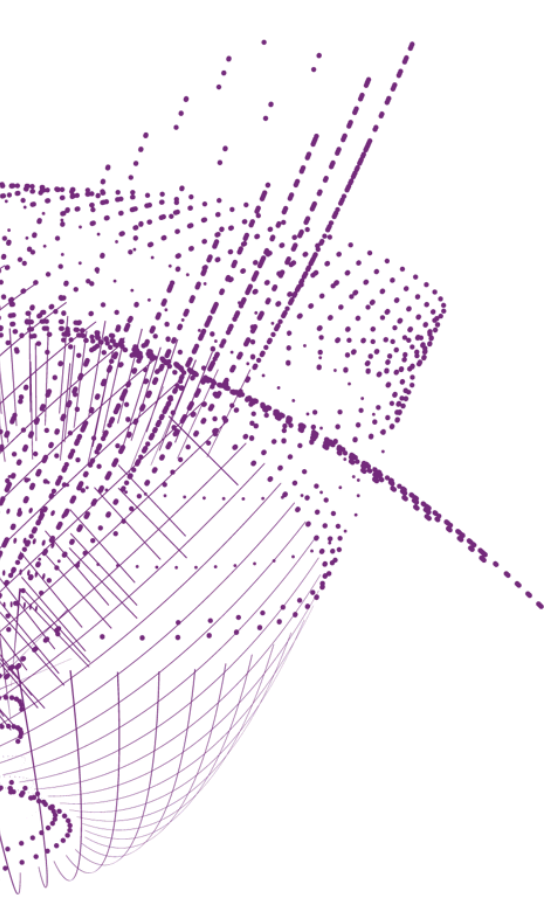
Data Management Plans

You have data, you have software, you have a PID: charge!

Nikhef

A data management plan is there to help you*

- where did I put that data file?
- where was the source data for this plot from?
- is the place where I write the results a safe one?
- do I understand what the columns in this file mean? Also next year?
- where did *my predecessor* put that data?
- What the $*&^$$%^& does the data in this directory mean?

The Data Management Plan "DMP" helps you **structure your data**, consider **proper formats**, ways to **annotate your data** (so you know what it means), … and how to make your **results outlive your laptop**.

* and your successor, your advisors, and colleagues!

Data Management Plan structure – just 6 questions

1. What data will be collected or produced, and what existing data will be re-used? (3 / 3) +

2. What metadata and documentation will accompany the data? (2 / 2) +

3. How will data and metadata be stored and backed up during the research? (2 / 2) +

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership? (2 / 2) +

5. How and when will data be shared and preserved for the long term? (6 / 6) +

6. Data management costs (1 / 1) +

NWO DMP format at https://dmponline.dcc.ac.uk/

When your data becomes part of something bigger

Data Management Plan elements

- What data will be collected or produced, and what existing data will be re-used?

  - Will you re-use existing data for this research?

  - If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

  - How much data storage will your project require in total?

When your data becomes part of something bigger

Useful data formats

Are commonly used, and
- should be self-describing
- allow metadata embedding
- be use re-use friendly, with a long (decades-long) comprehensibility
- OS, architecture & tool agnostic

Good examples
- root files
- CSV (HEPData compliant)
- HDF5 (and thus NetCDF)
- JSON – **with** a vocabulary

but for specific purposes, other formats are sometimes better

When your data becomes part of something bigger
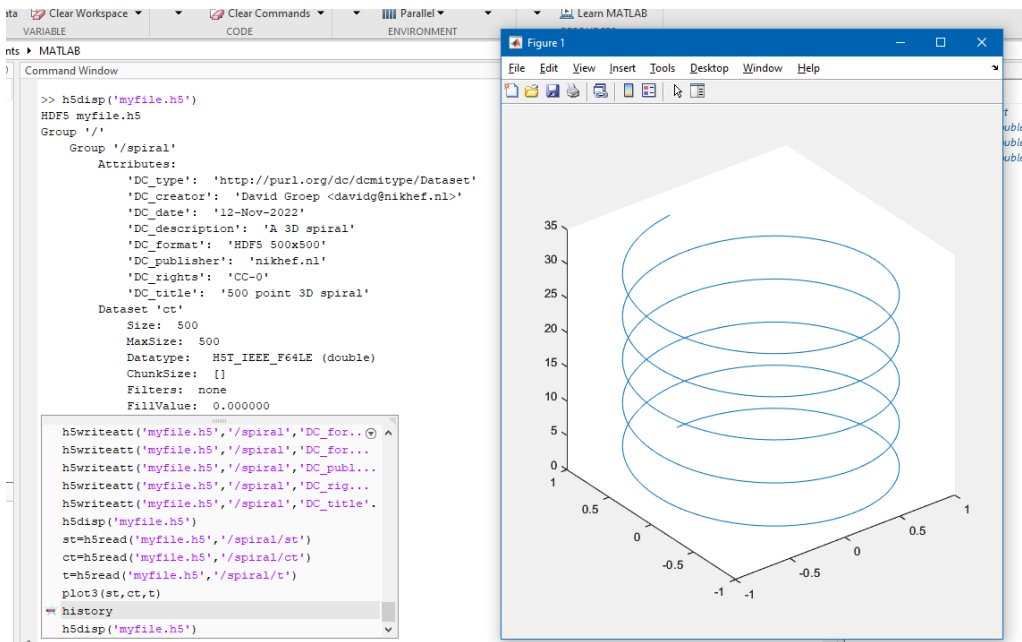
Organising data and meta-data

> - **What metadata and documentation will accompany the data?**
>   - Indicate what documentation will accompany the data
>   - Indicate which metadata will be provided to help others identify and discover the data.

Consider

- using a 'cookie cutter' to organize sources, references, results, https://cookiecutter.readthedocs.io/ (see Roel's session as well)
- and your repository/git convention, so add
  `README`, `LICENSE`, `NOTICE`, `CHANGES`, `CONTRIBUTING`
- *and then on to meta-data …*

As a cookie cutter example, see e.g. https://github.com/drivendata/cookiecutter-data-science

When your data becomes part of something bigger

Dublin Core and more

Already encountered the basic "Dublin Core" meta-data elements this morning

But you may need more **meta-data for interoperability**

- your experiment framework may provide that (alongside DC where useful)
- most of HEP is bespoke, given its long tradition and life span, but has meta-data … *we just never bothered to register, given the coherency of the discipline*
- **do ask** which other standards are relevant
  for example for GW, we also look at the IVOA standards
  https://www.ivoa.net/documents/RM/20070302/index.html
- and review the RDA meta-data standards catalogue https://rdamsc.bath.ac.uk/

most file formats support *embedding*: e.g. for Root objects, there's TTree::fUserInfo

https://root-forum.cern.ch/t/meta-data-in-root-files/16490/2 - see also the Research Data Alliance outputs at https://rd-alliance.org/

When your data becomes part of something bigger

When you still want to work with the data

- How will data and metadata be
  stored and backed up **during the research**?

  - Describe where the data and metadata will be stored and backed up during the project.

  - How will data security and protection of sensitive data be taken care of during the research?

When your data becomes part of something bigger

Put your data in the right place

During your work, use *managed systems*

- **avoid using only your local laptop** for storage,
  so use SURFdrive for syncing non-reproducible plots and docs

- external disks/USB thumbdrives are for **transfer only**

- publishable results, code, scripts, meta-data?
  these should be in /project/<groupname>

- bulk data, events, large data that can be generated
  should be in /data (small volumes) or /dcache

- data suitable for re-use: in your experiment DDM system, in Zenodo, arXiv, or SURFDataRepository

See https://www.nikhef.nl/pdp/doc/storage-classes

Summary: Data storage at Nikhef comes with
and care. For example, your home directory sh
up. Read about which type of files should go w

Table of Contents
- Home directory
- Data in /data
- dCache
- Project
- Local cache storage
- SURFdrive
- FileSender

When your data becomes part of something bigger

Some data needs more care than others

- **How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?**

  - Will you process and/or store personal data during your project?

  - How will ownership of the data and intellectual property rights to the data be managed?

Consider

- Do you collaborate with industry? What does the consortium agreement say?

- Do you use information 'under NDA' as an input? Review how it affects the results!

When your data becomes part of something bigger

When you are done with your data ... but the world isn't

- **How and when will data be shared and preserved for the long term?**

  - How will data be selected for long-term preservation?

  - Are there any (legal, IP, privacy related, security related)
    reasons to restrict access to the data once made publicly available, to
    limit which data will be made publicly available, or to not make part of the data publicly
    available?
  - What data will be made available for re-use?
  - When will the data be available for re-use, and for how long will the data be available?
  - In which repository will data be archived and made available for re-use, under which
    license?
  - Your strategy for publishing the analysis software that will be generated in this project?

When your data becomes part of something bigger

Nikhef

Long-term preservation – once you (think) you're done

'Curation may be the art of throwing away' – but what to keep?

keep things **'relevant for re-use'**

- obviously: all data that is used directly in publications

- data that can be needed to *reproduce the analysis* (also for research integrity)

- for a few specific things: there can be *regulatory requirements* to keep it

See e.g. https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data

When your data becomes part of something bigger

Not all data is, or should, be public

Quite obvious for personal data ("GDPR") – but we don't have much of that

But

- It may be '**dual-use**', and then subject to grant conditions or regulatory constraints

- It could disclose data we got originally as 'commercial in confidence' (under **NDA**)

- Even if not dual-use, also non-published results it may still be sensitive
  think of potential abuse-cases!

- any irking about what your data or project could cause? Talk about it!
  there is the 'loket Kennisveiligheid', and we have access to more sources

- Is your result patentable? Then you should file a patent *before* publishing

For knowledge safety, ethical, and espionage concerns, contact me, ronalds@nikhef.nl, or avr@nikhef.nl

Is Open Data always open *right now* ?

Data in *publications*, data points in plots, should be open alongside it,

For the rest
- data can be **embargoed** (be in a '**proprietary period**')
  for most LHC experiments 5 years after run ends, for LIGO 18 months, …
- for bulk data does not make sense: e.g. raw LHC data ('level 4') are not released
- think of 'non-intuitive' cases, e.g. some raw data for machine learning research, or how other domains can re-use data
- 'open by default, closed only when necessary' – and include the needed software

and in the data management plan, describe *what* data is made available *when*
''where' we already discussed: your collaboration open data system, or Zenodo, or …

See e.g. http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments , https://www.gw-openscience.org/, and https://dcc.ligo.org/LIGO-M1000066/public

When your data becomes part of something bigger

The million-euro question … literally ☺

- **Data management costs**

- What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable

---

Keep in mind

- *Open Access publication costs money, and*
- *then you store a lot of data, the repository must charge for it as well*

Storing data is quite costly:
if data is actually used, it is ~80 Eur/TB/year, and even for 'archival' data it is ~15 Eur/TB/year!
*The Repository - meta-data & anything beyond transfer accessibility – needs both storage **but also effort***

Filling your Data Management Plan

- a DMP is there to make your results 'FAIR', so be kind to your peers and do things reasonable for our work. So for GW, use formats that are discipline relevant (like IVOA meta-data), for LHC things use Root files.

  For other kinds of arrays (like geo data, detector measurements), use a structured self-describing format like HFD5 or a well-documented CSV.

- don't re-invent the wheel, there are Nikhef-specific examples!
- anyway, there is a Nikhef Data Management Policy … www.nikhef.nl/pdp/rdm/

- use an on-line DMP tool to guide you through the process

  Try it now as a mock DMP on https://dmponline.dcc.ac.uk/

https://www.nikhef.nl/pdp/rdm/

Examples for pure-LHC centred projects

https://www.nikhef.nl/pdp/doc/dmp-templates

When your data becomes part of something bigger

DMP Online Tool – a global tool

Also used by NWO: https://dmp.nwo.nl/ - it is the same instance

When your data becomes part of something bigger

Data reproducibility
since data management is 'just good science'

besides obvious issues you already know:
- results altered or omitted to make it 'look good'
- use open data to identify selection bias (or just 'interesting' use of statistics …)

less obvious:
confirmation bias – keep analysing until it fits
- *we mostly use blinding, but that's because we learnt …*
- *and why we plots results as a function of time!*
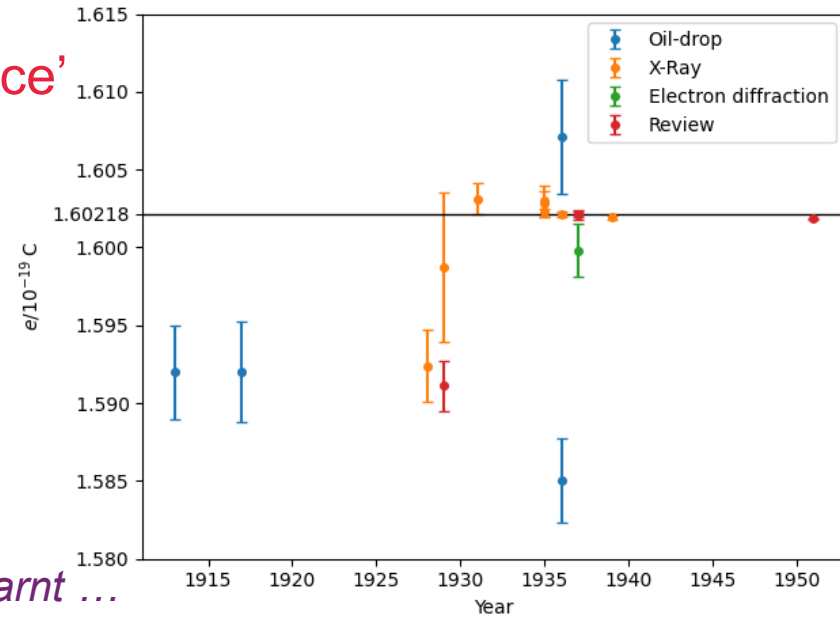
and some things only become apparent decades later …

Image from Christian Hill, https://scipython.com/blog/measurements-of-the-electron-charge-over-time/ (CC-BY-SA). See also Feynman, 1986
see also https://www.nwo-i.nl/en/employees/work-and-behaviour/scientific-integrity/

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

Maastricht University

Nikhef