

The logo for Nikhef, featuring the word "Nikhef" in a stylized, light blue font. The letter "i" is lowercase and has a vertical line through it. The letter "h" is lowercase and has a vertical line through it. The letter "e" is lowercase and has a vertical line through it. The letter "f" is lowercase and has a vertical line through it. The letters "N", "k", and "h" are uppercase. The letters "i", "e", and "f" are lowercase. The logo is set against a dark blue background.The logo for Maastricht University, featuring a stylized "U" and "M" in a dark blue triangle. The "U" is above the "M".

Maastricht University

Identifiers, repositories, licenses, and DMPs

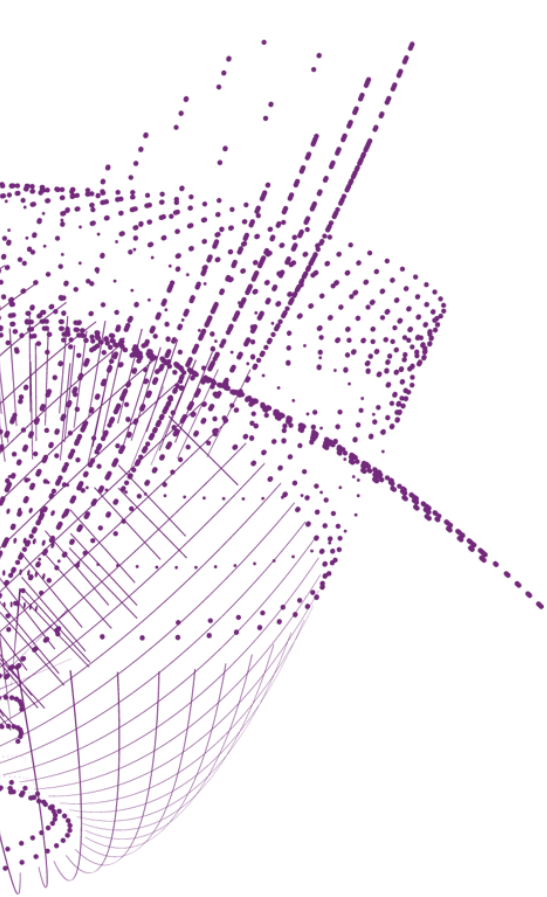
When your data
becomes part
of something bigger

An abstract graphic on the right side of the slide, consisting of a light blue background with a dark blue diagonal line. The graphic features a series of overlapping, curved lines and dots, creating a sense of depth and movement. The lines are in various shades of blue and purple, and the dots are in shades of blue and purple. The overall effect is a complex, layered pattern that suggests data or a network.

David Groep
Nikhef
Computing Course 2024

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!

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

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*

Not all identifiers are created equal


Technical qualities, but also ‘impact perception’

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

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

Uniform Resource Indicators == URLs + URNs

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



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:sproxy: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

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)
- HADRON Nikhef institutional repository (<https://doi.org/10.60953/uuid>)
- 4TU.RD (hosted at TU Delft)
- and commercial services like Figshare

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

When your data becomes part of something bigger



PHYSICAL REVIEW C
covering nuclear physics

Highlights Recent Accepted Collections Authors Referees Search Press About Edit

Investigation of the exclusive ${}^3\text{He}(e, e'pp)n$ reaction

D. L. Groep et al.
Phys. Rev. C **63**, 014005 – Published 19 December 2000

Article References Citing Articles (28) PDF Export Citation

ABSTRACT

Cross sections for the ${}^3\text{He}(e, e'pp)n$ reaction were measured over a wide range of energy and three-momentum transfer. At a momentum transfer $q = 375\text{ MeV}/c$, data were taken at transferred energies ω ranging from 170 to 290 MeV. At $\omega = 220\text{ MeV}$, measurements were performed at three q values (305, 375, and 445 MeV/c). The results are presented as a function of the neutron momentum in the final state, as a function of the energy and momentum transfer, and as a function of the relative momentum of the two-proton system. The data at neutron momenta below 100 MeV/c, obtained for two values of the momentum transfer at $\omega = 220\text{ MeV}$, are well described by the results of continuum-Faddeev calculations. These calculations indicate that the cross section in this domain is dominated by direct two-proton emission induced by a one-body hadronic current. Cross section distributions determined as a function of the relative momentum of the two protons are fairly well reproduced by continuum-Faddeev calculations based on various realistic nucleon-nucleon potential models. At higher neutron momentum and at higher energy transfer, deviations between data and calculations are observed that may be due to contributions of isobar currents.

Received 9 August 2000

DOI: <https://doi.org/10.1103/PhysRevC.63.014005>

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>

GW150914

Version 1 Version 2 **Version 3**

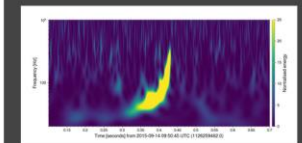
Documentation

Release: GWTC-1-confident
Event UID: GW150914-v3
Names: GW150914
GPS: 1126259462.4
UTC Time: 2015-09-14 09:50
Timeline: Query for segments
DOI: <https://doi.org/10.7935/82H3-HH23>

<https://doi.org/10.7935/82H3-HH23>

Event from GWTC-1. For documentation, see:
<https://arxiv.org/abs/1611.12907>
<https://doi.org/10.7935/82H3-HH23>

H1 strain



32sec - 16KHz:	GWF	HDF	TXT
32sec - 4KHz:	GWF	HDF	TXT
4096sec - 16KHz:	GWF	HDF	TXT
4096sec - 4KHz:	GWF	HDF	TXT

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:Z $\tau\tau$ + Jets, for 2016 ATLAS open data release

Z $\tau\tau$ + Jets, ATLAS Collaboration

Cite as: ATLAS Collaboration (2016). MC:Z $\tau\tau$ + 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.



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

DOIs can be assigned to almost any object

The screenshot shows the Zenodo interface for a document titled "FIM4R Position Paper on the Desired Evolution of EOSC Authentication and Authorisation Infrastructures". The document is dated March 26, 2020, and is categorized as a "Working paper" and "Open Access".

The "Basic information" section on the right includes a "Digital Object Identifier" field with the example "e.g. 10.1234/foo.bar" and a "Reserve DOI" button. Below this, it states: "Optional. Did your publisher already assign a DOI to your upload? others to easily and unambiguously cite your upload. Please note it is always possible to edit a custom DOI."

The document is indexed in "OpenAIRE". The "Publication date" is March 26, 2020, and the "DOI" is 10.5281/zenodo.3727546, which is highlighted by a red arrow. The "Keyword(s)" are EOSC, FIM4R, and AAI. The "Communities" listed are Federated Identity Management for Research. The "License (for files)" is Creative Commons Attribution 4.0 International.

A preview of the document is shown on the left, with the title "FIM4R Position Paper" and subtitle "On the Desired Evolution of EOSC Authentication and Authorisation".

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



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:

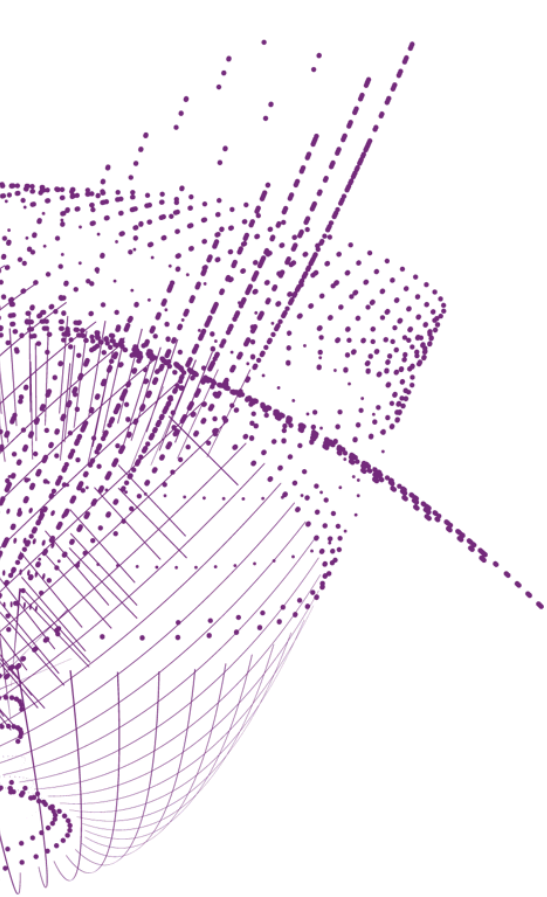
4b. Key output

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.

- for papers, journal will assign one
- arXiv links are persistent as well and link to later DOIs

11. arXiv:nucl-ex/0703007 [pdf, ps, other]   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

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



Identifying you ...

ORCID – the Open Researcher 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

Consortium
Please list in the table below all consortium members. The call distinguished four different categories of consortium members: 1) main applicant; 2) co-applicant(s); 3) co-funder(s) (not compulsory); 4) cooperation partner(s) (not compulsory). Adhere to these categories in the table below. Remove categories 3 and/or 4 if not relevant. The main applicant and co-applicants (categories 1 and 2) must meet the requirements for (co-)applicants as stated in the call for proposals, Section 3.1.1, and must also be entered in ISAAC.

Title, first name, initials, surname	Affil.	Expertise	ORCID/website/DAI/other	Role(s)
Main applicant				
Prof Dr E. (Edward) Example	TUD RUG	Queueing theory	orcid.org/0000-0000-0000-0000	WP3 leader
Co-applicant(s)				
Dr V. (Vera) Vorschlag	RUG	Peer-to-peer networks, Data science	www.vorschlag.com	WP2 contributor
V. (Victor) Voorbeeld	Philips	Cooling liquids	www.phillips.com/groupvb	Technical support

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>

Recognising and using ORCID with linking

- based on the ISNI format

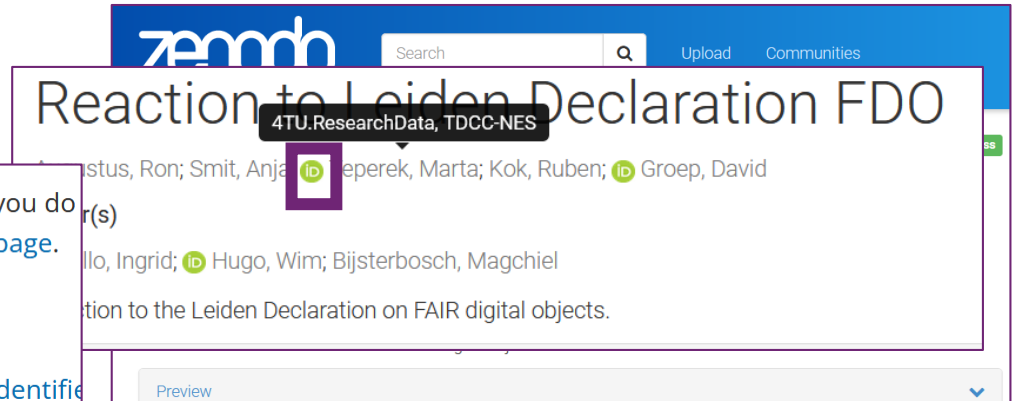
orcid.org/**0000-0003-1026-6606**

- can be linked automatically in journals and repositories

We encourage all arXiv authors to link their ORCID iD with arXiv. If you do so, it will be used in the process. Once completed you will see your ORCID iD on your [user page](#).

 [Link my arXiv account with ORCID](#)

arXiv will use ORCID iDs in preference to the internal [arXiv author identifier](#)



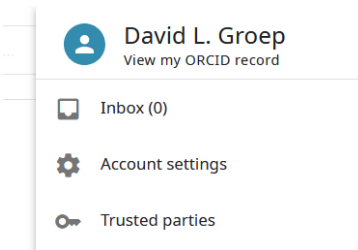
The screenshot shows a Zenodo record page. The title is "Reaction to Leiden Declaration FDO". Below the title, there is a list of authors: "Stus, Ron; Smit, Anja;  eperek, Marta; Kok, Ruben;  Groep, David". A tooltip points to the ORCID icon for "eperek, Marta" with the text "4TU.ResearchData, TDCC-NES". Below the author list, there is a "Preview" button.

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

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*)



▼ 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.](#)

Organization's name institution logo

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

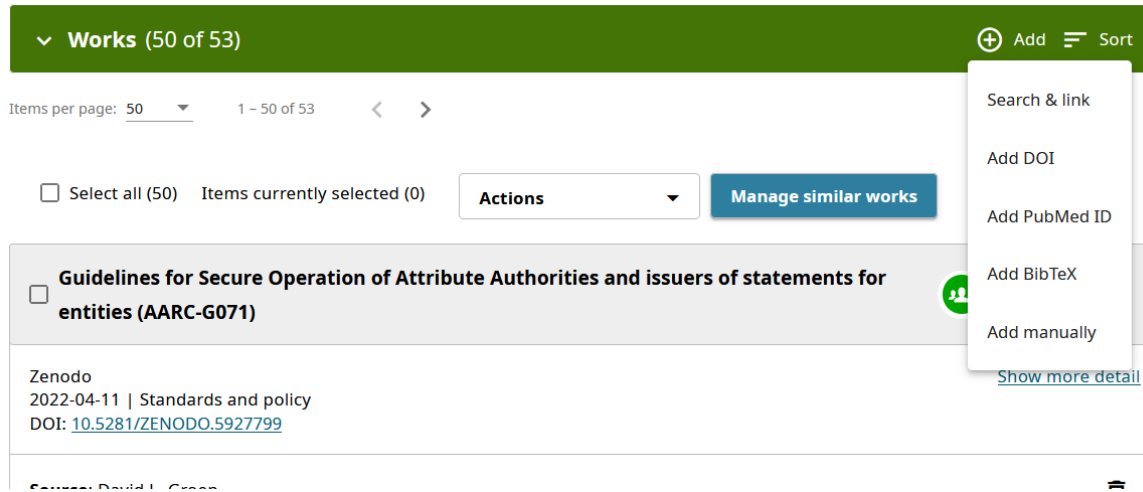
You can sign into ORCID using the personal and institutional accounts you have linked to your ORCID record.

[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	

Linking your publications to ORCID

Import from Scopus and commercial publishers



The screenshot shows the Zenodo interface for a list of works. At the top, there is a green header with a dropdown arrow and the text "Works (50 of 53)". To the right of the header are icons for "Add" (a plus sign) and "Sort" (three horizontal lines). Below the header, there is a section for "Items per page: 50" and "1 - 50 of 53" with navigation arrows. A checkbox labeled "Select all (50)" is followed by "Items currently selected (0)". To the right of this is a dropdown menu labeled "Actions" and a blue button labeled "Manage similar works". The main content area shows a list of works. The first work is "Guidelines for Secure Operation of Attribute Authorities and issuers of statements for entities (AARC-G071)". Below the title, it shows "Zenodo", "2022-04-11 | Standards and policy", and "DOI: 10.5281/ZENODO.5927799". A context menu is open over the work, showing options: "Search & link", "Add DOI", "Add PubMed ID", "Add BibTeX", "Add manually", and "Show more detail".

logging in to Zenodo with your ORCID

Some agencies can *update* your record for you

You do get a notice, like this:

and you can associate your ORCID with your Nikhef account (can be shared with services, but not much used yet)

address	
Common name	David Groep
Unix GECOS (finger) data	David Groep,Nikhef Amsterdam H2.29b,+31205922179,+31646812179
ORCID	associate your ORCID
Contact address	Room: H229b Telephone number: +31205922179
Picture visible in public gallery	Yes (change this setting here)

<https://sso.nikhef.nl/myorcid/>

ORCID Authorize access

You are currently signed in as:

David L. Groep
<https://orcid.org/0000-0003-1026-6606>
[Sign out](#)

[OpenAIRE Explore ?](#)

has asked for the following access to your ORCID record:

- Add/update your research activities (works, affiliations, etc)
- Read your information with visibility set to Trusted Organizations

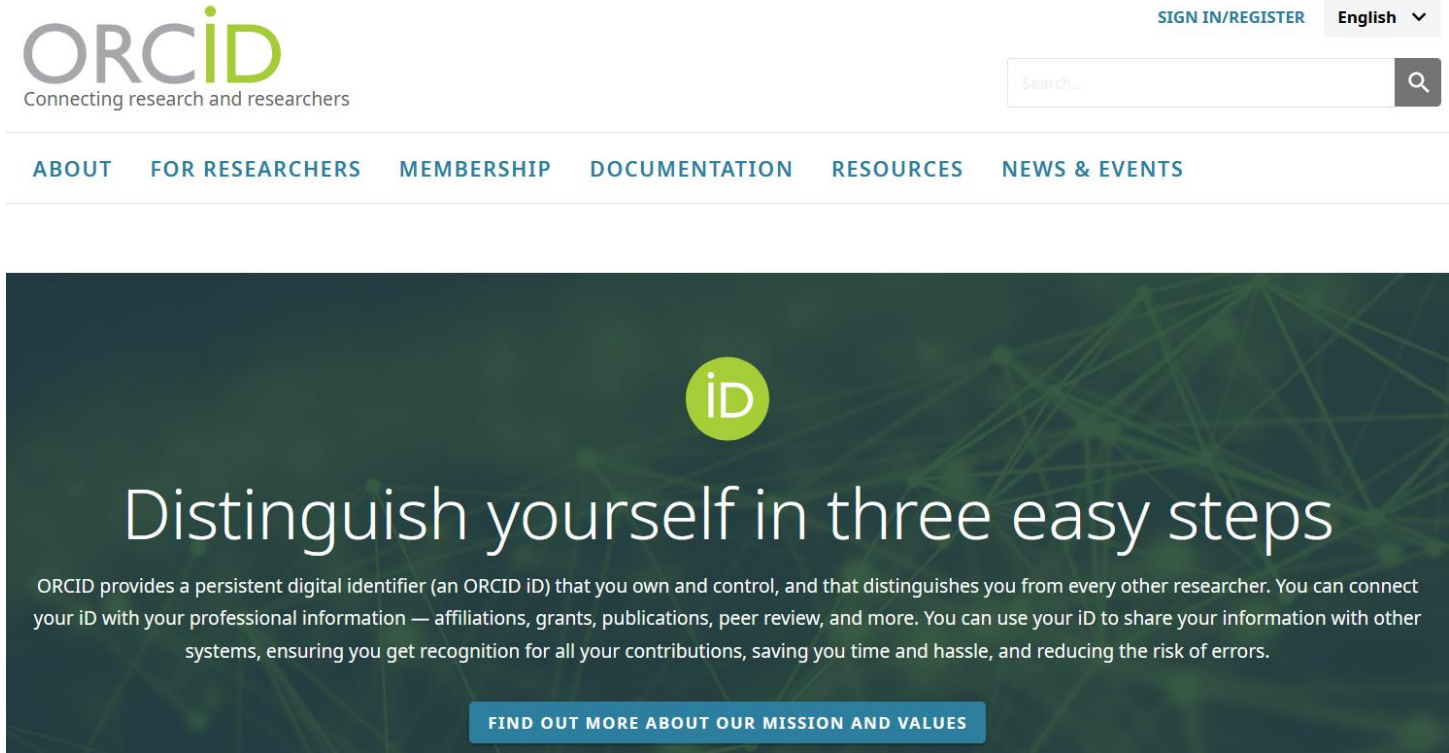
[Authorize access](#)

[Deny access](#)

If authorized, this organization will have access to your ORCID record, as outlined above and described in further detail in [ORCID's privacy policy](#).

You can manage access permissions for this and other Trusted Organizations in your [account settings](#).

Now go over to orcid.org, and ...



The screenshot shows the ORCID website homepage. At the top left is the ORCID logo with the tagline "Connecting research and researchers". To the right are links for "SIGN IN/REGISTER" and a language dropdown set to "English". Below this is a search bar with a magnifying glass icon. A horizontal navigation menu contains the following items: ABOUT, FOR RESEARCHERS, MEMBERSHIP, DOCUMENTATION, RESOURCES, and NEWS & EVENTS. The main content area features a dark background with a network diagram. It includes the ORCID iD logo (a green circle with "iD" in white), the headline "Distinguish yourself in three easy steps", and a paragraph explaining that ORCID provides a persistent digital identifier that distinguishes researchers and connects their professional information. At the bottom of this section is a blue button that says "FIND OUT MORE ABOUT OUR MISSION AND VALUES".

ORCID
Connecting research and researchers

SIGN IN/REGISTER English

Search...

ABOUT FOR RESEARCHERS MEMBERSHIP DOCUMENTATION RESOURCES NEWS & EVENTS

iD

Distinguish yourself in three easy steps

ORCID provides a persistent digital identifier (an ORCID iD) that you own and control, and that distinguishes you from every other researcher. You can connect your iD with your professional information — affiliations, grants, publications, peer review, and more. You can use your iD to share your information with other systems, ensuring you get recognition for all your contributions, saving you time and hassle, and reducing the risk of errors.

FIND OUT MORE ABOUT OUR MISSION AND VALUES

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



The image shows a screenshot of the Zenodo website's sign-up page. The background is a solid blue color. At the top center, the word "zenodo" is written in a white, lowercase, sans-serif font. Below the logo, the text "Research. Shared! Sign up today." is centered in a white, sans-serif font. A thin white horizontal line separates the header from the main content area. On the left side, there are three sections of text in white: "Citeable. Discoverable." followed by a paragraph about Digital Object Identifiers (DOIs); "Communities" followed by a paragraph about accepting or rejecting uploads; and "Trusted Research Data Management" followed by a paragraph about CERN's expertise. On the right side, there are three white buttons stacked vertically: "Sign up with GitHub" (with a GitHub icon), "Sign up with ORCID" (with an ORCID icon), and a separator "— OR —". Below these buttons are three white input fields stacked vertically, labeled "Email Address", "Username", and "Password".

But what about PID and software?

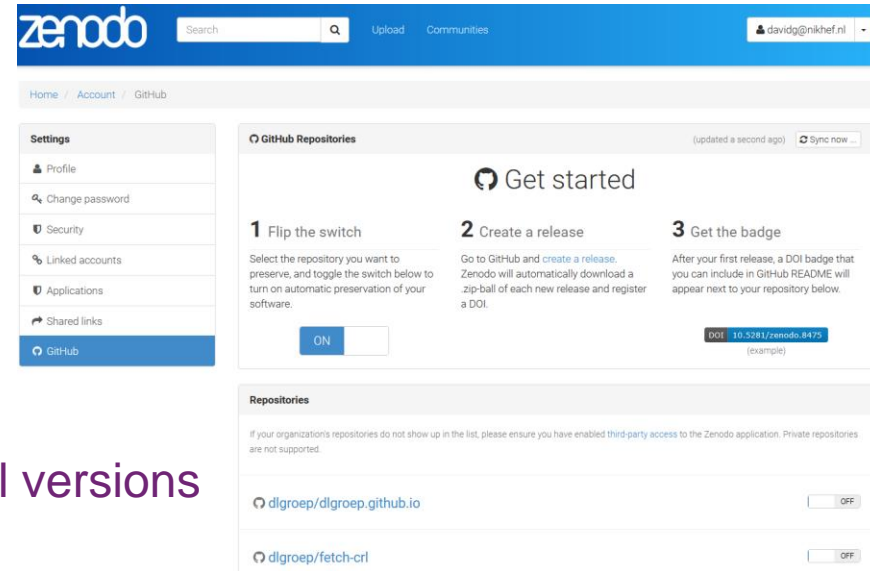
A personal github repository is not a persistent identifier either

- it is good for FOSS and collaboration, but it is not a publication ... *yet*
- from Zenodo, link your github account
- in Zenodo select your repository
- in Github, make a *release*

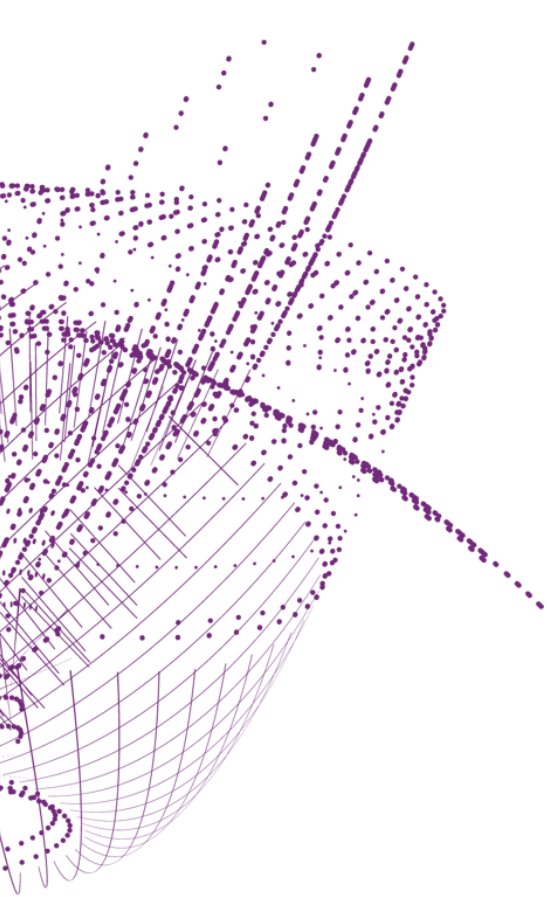
this will be automatically uploaded into Zenodo – and you have persistent DOIs for all versions

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>



The screenshot shows the Zenodo account settings page for GitHub integration. The page is titled "GitHub Repositories" and includes a "Get started" section with three steps: 1. Flip the switch, 2. Create a release, and 3. Get the badge. A toggle switch is currently in the "ON" position. Below the "Get started" section, there is a "Repositories" section with a list of repositories. The first repository is "dlgroep/dlgroep.github.io" and the second is "dlgroep/fetch-ctrl". Both repositories have a toggle switch in the "OFF" position. The page also includes a search bar, a user profile icon, and a "Sync now" button.



Licensing and open source

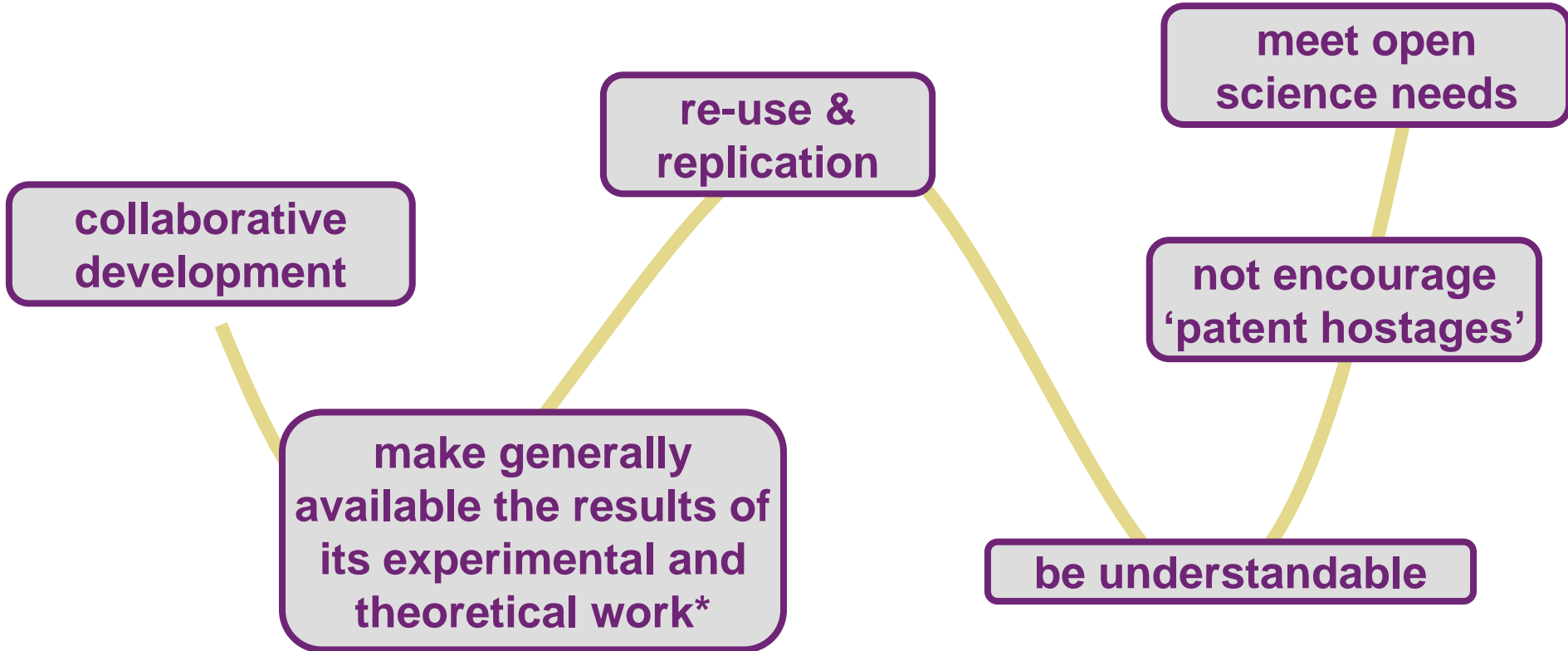
Enabling Software Re-use



however: code and source files without a license means ‘all rights reserved’



What do we actually want to achieve with our software?



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

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

Mozilla family
(MPL, Perl Artistic)

re-use friendly

patent protection

contributions unspecified

lesser copyleft
(LGPL)

re-use unfriendly

no patent protection

contributions auto-imported: N/A

Service transparent
(AGPL)

re-use unfriendly

3.0+patent protection

contributions auto-imported: N/A

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>

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”

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.

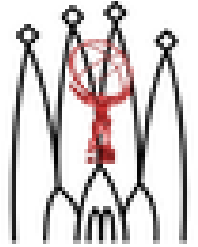
Accordingly, the LHCb software stack also needs to be licensed under the

“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

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*.

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 rights reserved.

- 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 ...

The Geant4 Software License

Established 30 June 2006 for Geant4 release 6.1, subsequent patches and releases.
Previous releases are covered by the disclaimer included in the release.

Copyright Holders of the Geant4 Collaboration

Last revision: 30 June 2006
The collaboration has established the following list of institutions and individuals who hold copyright of parts of

Institutions

Bath University, Bath, UK
Bayer Institute Nuclear Physics, Novosibirsk, Russia
Budapest Technical University, Budapest, Hungary
California Institute of Technology, Pasadena, USA
CERN, European Organization for Nuclear Research, Geneva, Switzerland
COMAT, Madrid, Spain
CURSI-IGP2, Institut National de Physique Nucléaire, France
ESA, European Space Agency
ETH, Zurich, Switzerland
Fermi National Accelerator Laboratory, Batavia, USA
Hawaii Institute of Physics, Hawaii, France
IHEP, Protvino, Russia
Imperial College, London, UK
Institut für Experimentelle Kernphysik, Karlsruhe University, Karlsruhe, Germany
Instituto de Física de Cantabria, Santander, Spain
INFN, Istituto Nazionale di Fisica Nucleare, Italy
IST National Institute for Cancer Research, Italy
Jefferson Laboratory, USA
JINR, Dubna, Russia
J. W. Goethe-Universität, Frankfurt am Main, Germany
Karolinska Institutet, Stockholm, Sweden
KFKI Research Institute for Particle and Nuclear Physics, Budapest, Hungary
Laboratório de Instrumentação e Física Experimental de Partículas (LIP), Lisbon, Portugal
Laboratoire, Moscow, Russia
Manchester University, Manchester, UK
MIT Massachusetts Institute of Technology, Cambridge, USA
Moscow Engineering Physics Institute (State University), Moscow, Russia
Northwestern University, Evanston, USA
Pittsburgh University, Pittsburgh, USA
Rutherford Appleton Laboratory, UK
Southampton University, Southampton, UK
Stanford University (SLAC, SLAC National Accelerator Laboratory), Stanford, USA
Tampere University, Tampere, Finland
TRIUMF, Vancouver, Canada
University of British Columbia, Vancouver, Canada
University of California, Santa-Cruz, USA
University of Cordoba, Cordoba, Spain
University of Maryland, USA
Centre For Medical Radiation Physics (CMRP), University of Wollongong, Australia

Individuals

Akron Koruna
Bil Szelei
Glenys Nelson-Smith
Harm Fesefeldt
Hajdu Kurtzke
Hajdu Tamas
Kalyay Martina
Kalyay Aniko
Kopov Viktor
Makoto Takahashi
Makoto Abe
Miklos Horvath
Marcus H. Mendenhall
Karl Wimmer
Robert A. Weber
Shang Se
Sebastian Tamas
Yoshiko Senada
Toshiaki Kodama
Takaya Yuzuru
Tuzsanyi Andras
Tamas Szeki
Yousef Morita

... 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

EGEE II started on 1 April 2006 and the new EGEE website can be found at: <http://www.eu-egee.org>

EGEE Partners

The EGEE Partners are those people and institutions that are currently using the Grid or providing a computational resource to it. The EGEE project consists of both contracting and non-contracting partners.

EGEE contracting partners have signed the EGEE contract and receive contributions from the EU, whereas non-contracting partners do not receive any EU contribution but are interested in the programme of work and participate in some EGEE activities.

A list of EGEE **non-contracting** partners is available here, and a list of Non-Contracting Partners who have signed a Memorandum of Understanding is available here.

Contracting Partners

#	Organisation	Acronym	Federation	Country
1	European Organization for Particle Physics	CERN	CERN	Switzerland
2	Institut für Graphische und Parallele Datenverarbeitung der Joh. Kepler Universität Linz	GUP	Central Europe	Austria
3	Institut für Informatik der Universität Innsbruck	UNIINNSBRUCK	Central Europe	Austria
4	CESNET, z.s.p.o.	CESNET	Central Europe	Czech Republic
5	Budapest University of Technology and Economics	BUTE	Central Europe	Hungary
6	Eotvos Lorand University Budapest	ELUB	Central Europe	Hungary
7	KFKI Research Institute for Particle and Nuclear Physics	KFKI RMKI	Central Europe	Hungary
8	Magyar Tudományos Akademia Szamiastecnikai es Automatalizalasi Kutato Intezet	MTA SZTAKI	Central Europe	Hungary
9	Office for National Information and Infrastructure Development	NIFI	Central Europe	Hungary
10	Akademickie Centrum Komputerowe CYFRONET akademii Gorniczo-Hutniczej im.St. Staszica w Krakowie	CYFRONET	Central Europe	Poland
11	Warsaw University Interdisciplinary Centre for Mathematical and Computational Modelling	ICM	Central Europe	Poland

Sources:
from the GEANT4 web pages at
<https://geant4.web.cern.ch/license>
and
<http://eu-egee.org/partners>

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,
```

e.g. <https://github.com/filesender/filesender/>

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

Reinder's NWO-I LDCC License Tool:
<https://www.nikhef.nl/pdp/rdm/license-tool>

Nikhef RDM and Licensing guidelines:
<https://www.nikhef.nl/pdp/rdm/policies>



from: <https://opensource.org/licenses>

A last word about patents ...

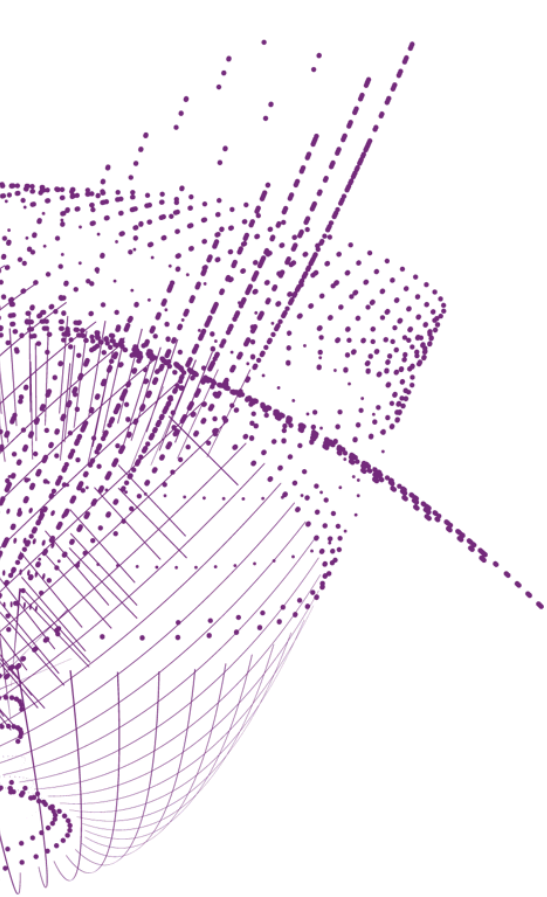
... usually for mutual litigation, but have been used against open source (although rejected)

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 Ig Nobel prize, of course!



Data Management Plans

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

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/>

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?

Useful data formats

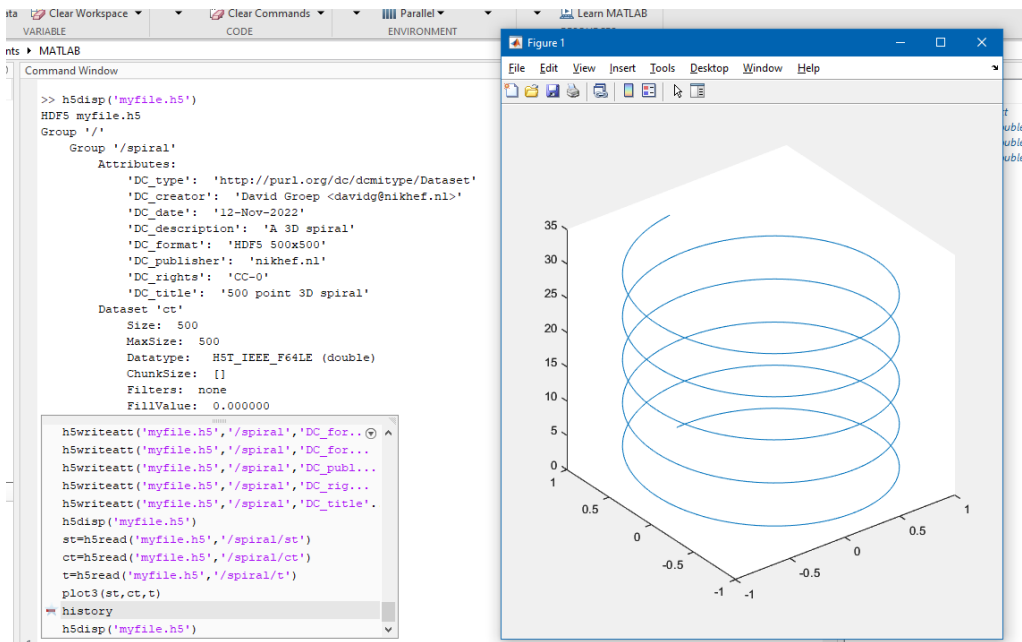
Are commonly used, and

- 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



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>

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 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?

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://kb.nikhef.nl/ct/Storage_overview.html

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

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 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?

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>

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>

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

Services and software



- About the NDPF
- News and events
- Services and Resources
- Computing course
- Service documentation
- Research Data Management**
- Nikhef DMP Policies
- DMP templates
- License Selection Tool
- LDCC Digital Competence Centre
- Other services
- Systems
- Software and Tools

Data management templates

Summary: These documents provide good background and some copy-paste text that can be used to generate or fill the required NWO Data Management Plan (DMP) documents for any project related to the LHC experiments. The text you can copy from here, actually making a properly-formatted DMP is most easily done in [DMP Online] (<https://dmp.nwo.nl/>).

Note: Really - use the [DMPOnline tool](#) and save everyone a lot of work.

Forms for use in NWO submission

 Formulier Datamanagementplan ENG - ATLAS-v03-plan.docx	Nikhef template LHC ATLAS DMP form
 Formulier Datamanagementplan ENG - ATLAS-v03-plan.pdf	Nikhef template LHC ATLAS DMP form
 Formulier Datamanagementplan ENG - Alice-v01-	Nikhef template LHC Alice DMP form

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

DMP Online Tool – a global tool

The image displays two screenshots of the DMP Online Tool interface. The left screenshot shows a project plan titled "Physical laws: from pandemics to black holes" with the "Write Plan" tab selected. The right screenshot shows the "Create a new plan" form, where a red box highlights the checkbox for "mock project for testing, practice, or educational purposes".

Left Screenshot: Project Plan Overview

- Project Details | Contributors | Plan overview | **Write Plan** | Share
- expand all | collapse all | 19/18 a
- General Information (2 / 2)
- 1. What data will be collected or produced, and what existing data will be used? (2 / 2)
- 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? (2 / 2)
- 5. How and when will data be shared and preserved for the long term? (2 / 2)
- 6. Data management costs (1 / 1)
- 6.1 What resources (for example financial and time) will be dedicated to the management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?**

Right Screenshot: Create a new plan

Before you get started, we need some information about your research project to set you up with the best DMP template for your needs.

* What research project are you planning?
Electron-induced two-nucleon knock-out from 3He at 564 MeV
If applying for funding, state the project title exactly as in the proposal.

mock project for testing, practice, or educational purposes

* Select the primary research organisation
Organisation: - or -
Begin typing to see a list of suggestions.

* Select the primary funding organisation
Funder: - or -
NWO
Netherlands Organisation for Scientific Research (NWO)

<https://dmponline.dcc.ac.uk/plans/new> - you can login with your Nikhef or university credentials

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 ...

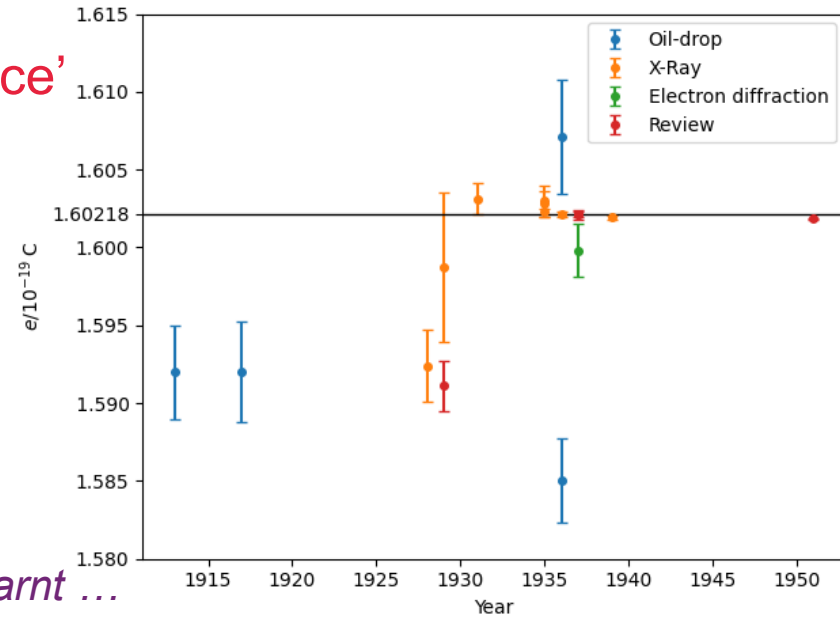


Image from Christian Hill, <https://scipython.com/blog/measurements-of-the-electron-charge-over-time/> (CC-BY-SA). See also Feynman, 1986 see also <https://www.nwo-i.nl/en/employees/work-and-behaviour/scientific-integrity/>



Maastricht University

Nikhef

David Groep

davidg@nikhef.nl

<https://www.nikhef.nl/~davidg/presentations/>

 <https://orcid.org/0000-0003-1026-6606>

