

“Module data sharing, archiving and publishing”

An introduction....

Cees H.J. Hof

Data Archiving and Networked Services (DANS-KNAW)

With the help of Jasmin Böhmer (UMCU) & Christine Staiger (DTL)

 @CeesH_DANS

29 MAY 2019, Helis Academy, FAIR Data Stewardship Course

Darwin Incubator, Niel, Belgium

About your trainer.....

- At DANS:
 - Project acquisition
 - Liaison life sciences
 - European Open Science Cloud (EOSC)
 - Software sustainability
 - Coach “Essentials 4 Data Support”
- +10 years involved in development of Global Biodiversity Information System (GBIF)
 - FAIR data *avant la lettre*
 - Cataloguing biodiversity data
 - Developing and implementing the DarwinCore data standard
 - Community building
- Background in Biology (ecology, taxonomy & paleontology)



GBIF

www.gbif.org

DANS is about keeping data FAIR



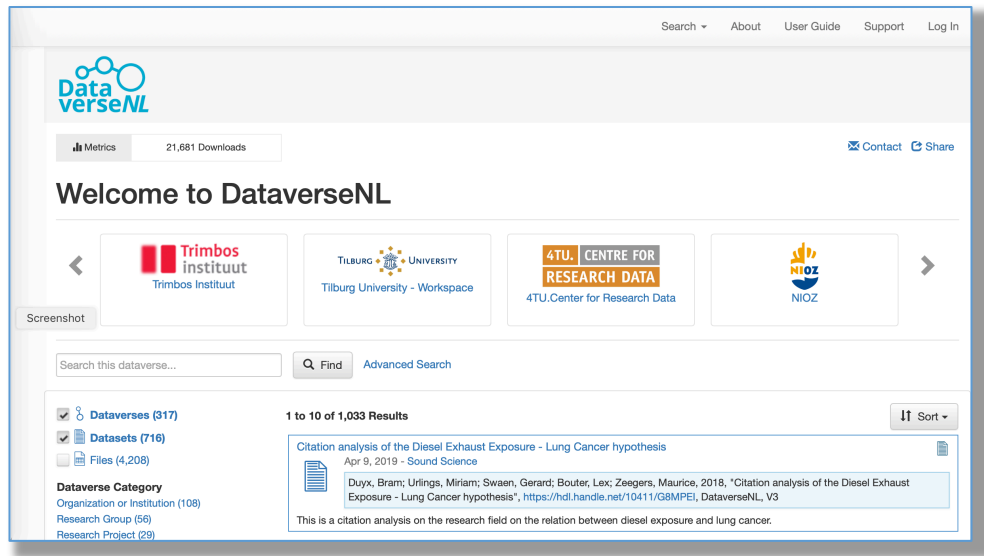
Mission: promote
and provide
permanent
access to digital
research
resources

Institute of
Dutch Academy
and Research
Funding
Organisation
(KNAW & NWO)
since 2005

First predecessor
dates back to
1964 (Steinmetz
Foundation),
Historical Data
Archive 1989

About DANS core services

DataverseNL



The screenshot shows the DataverseNL website interface. At the top, there is a navigation bar with links for Search, About, User Guide, Support, and Log In. Below the navigation bar, the website logo "DataverseNL" is displayed. A metrics section shows "21,681 Downloads" and links for Contact and Share. The main heading is "Welcome to DataverseNL". Below this, there are logos for participating organizations: Trimbos Instituut, Tilburg University - Workspace, 4TU. Centre for Research Data, and NIOZ. A search bar is present with the text "Search this dataverse..." and buttons for Find and Advanced Search. The search results section shows "1 to 10 of 1,033 Results". The first result is a citation analysis titled "Citation analysis of the Diesel Exhaust Exposure - Lung Cancer hypothesis" dated Apr 9, 2019 - Sound Science. The citation text is: "Duyx, Bram; Urlings, Miriam; Swaen, Gerard; Bouter, Lex; Zeeegers, Maurice, 2018, 'Citation analysis of the Diesel Exhaust Exposure - Lung Cancer hypothesis', <https://hdl.handle.net/10411/GBMPEI>, DataverseNL, V3". Below the citation text, there is a note: "This is a citation analysis on the research field on the relation between diesel exposure and lung cancer."

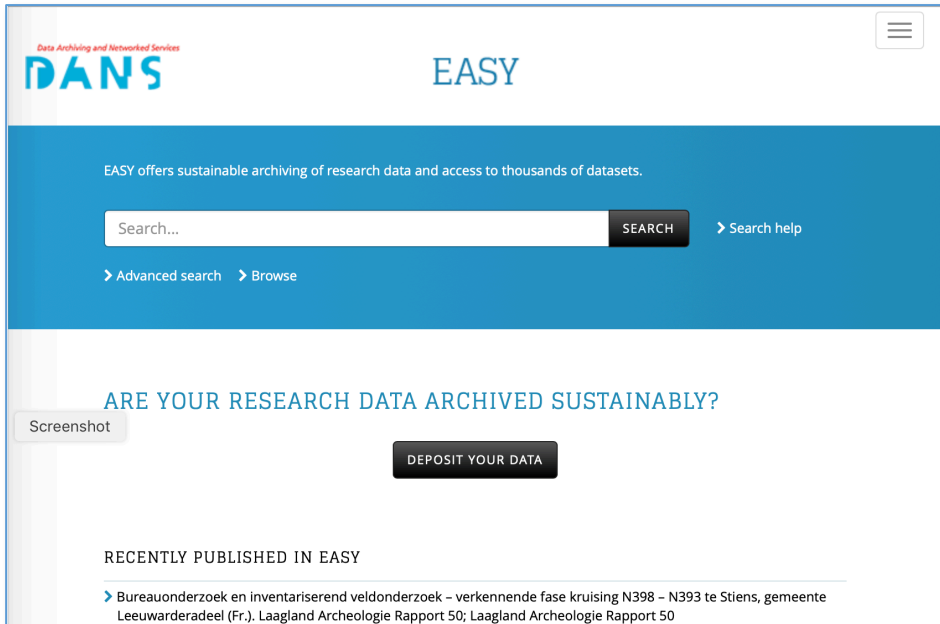
For RDM in ongoing research

- Including short / intermediate term storage
- Discipline agnostic
- 14 organisations participating
- DANS only technical maintenance and support
- Ongoing development

<https://dataverse.nl>

About DANS core services

EASY



For long term storage

- **Certified according to Core Trust Seal**
- **Discipline agnostic**
- **Biased towards humanities – social sciences – life sciences**
- **> 80.000 datasets**
- **Team of data custodians**
- **Below 1 TB free service**
- **Ongoing developments**

<https://easy.dans.knaw.nl/>

About DANS core services

NARCIS



<https://www.narcis.nl/>

Gateway to scholarly information in the Netherlands

- **Metadata only**
- **Around 2 million publications**
- **More than 200.000 datasets**
- **Tool to monitor research output**
- **Harvesting info from local crisis databases**
- **Using OAI-PMH protocol**
- **DublinCore and DataCite as main standards**
- **Providing metadata to European and international networks**

RDM training by DANS



In the context of many H2020 projects...



Browse through our recent webinars

Joint webinar FREYA and OpenAIRE: New developments in the field of Persistent Identifiers

The importance of Persistent Identifiers (PIDs) to build stable connections between research entities such as grants, projects, articles, or funders is recognized and addressed by several initiatives and projects.

Thursday, 10 January 2019

At the national level...

Objectives of this module:

Basic knowledge of:

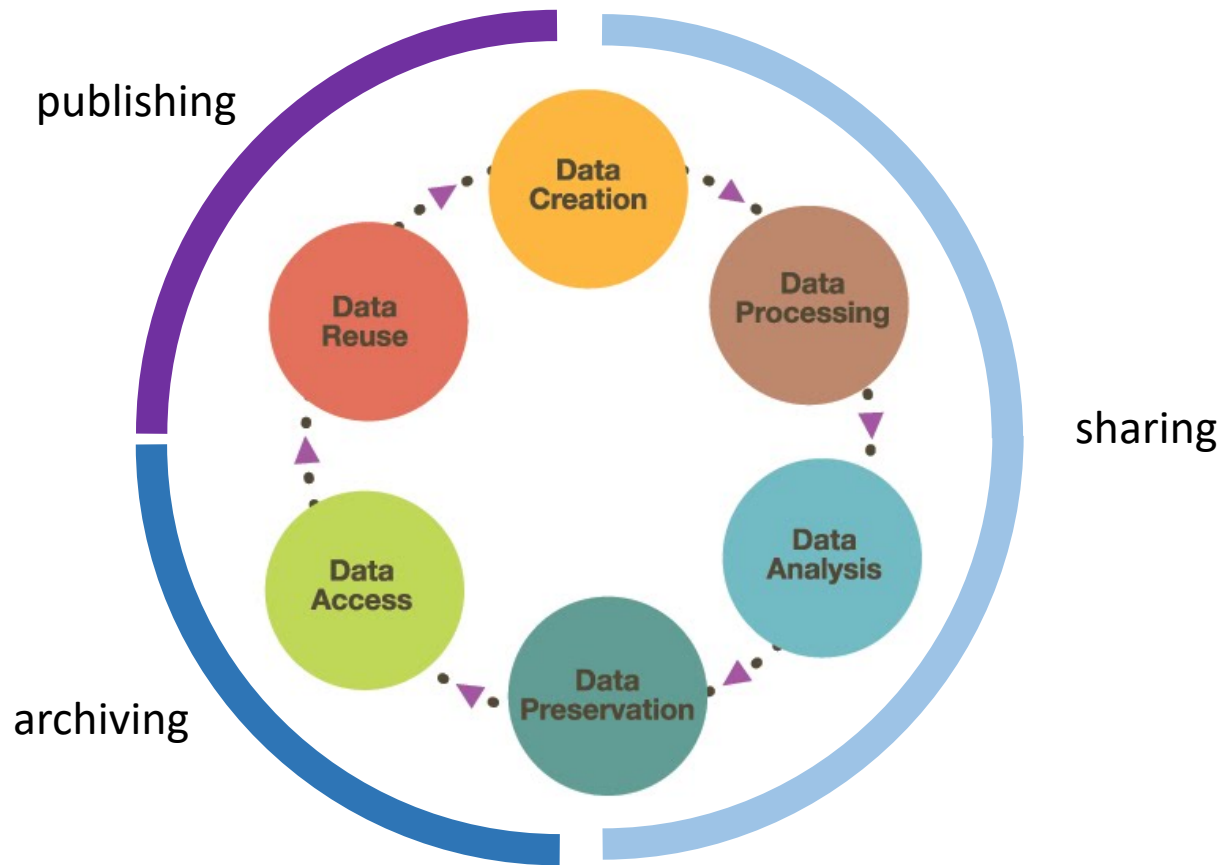
- Know what a repository is
- Characteristics of Data Sharing, Archiving and Publishing
- Knowledge of major platforms for Sharing, Archiving and Publishing
-

Hands on experience:

- Dataverse possibilities for sharing and publishing
- Designing requirements digital archives
- Finding your ideal archive
- A bit of FAIR assessment..... (if time allows)

Terminology....

Sharing versus archiving versus publishing...



Sharing



Tools to store and share data

- Dropbox, Google Drive, GitHub (commercial), DataVerse, Figshare
- B2SHARE (EUDAT)
- Owncloud-based services: SURFdrive, Research Drive (national NL)
- A-Z drives for Windows (institutional)

Characteristics:

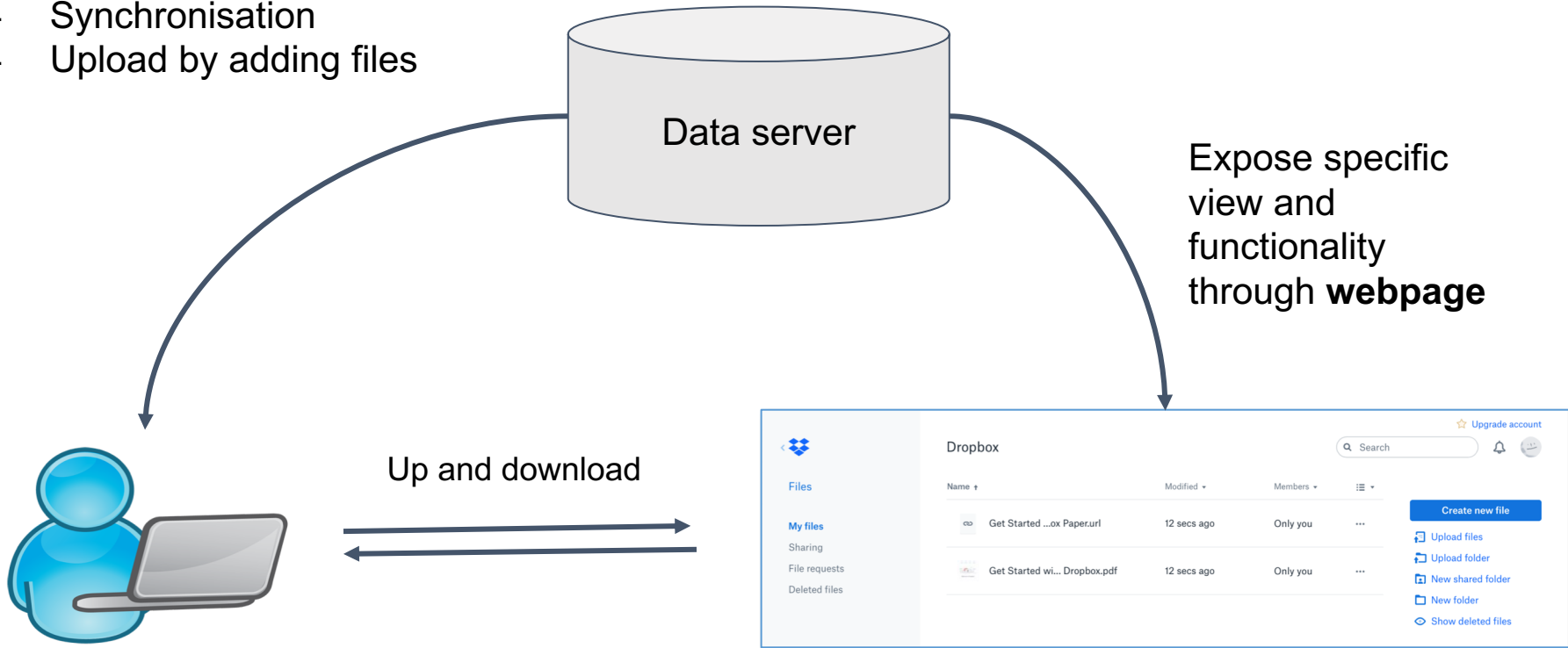
- Storing active data
- Data can be easily changed
- Some tools do have a versioning policy
- Data sharing main purpose

Sharing

how do these services work?

Access by filesystem Mounts:

- Data lies on server
- Extra folder on computer
- Synchronisation
- Upload by adding files



Sharing

- Data owner (researcher) has full command
 - Decides with whom to share
 - Can revoke access
 - Can delete data

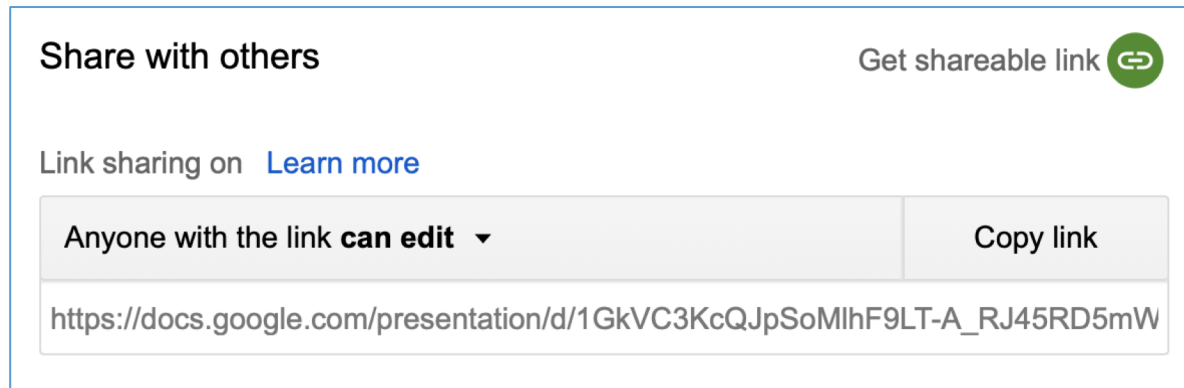
→ Huge flexibility


→ No extra security through third person, **risk!**

Sharing

Ways of sharing data

-Shareable link



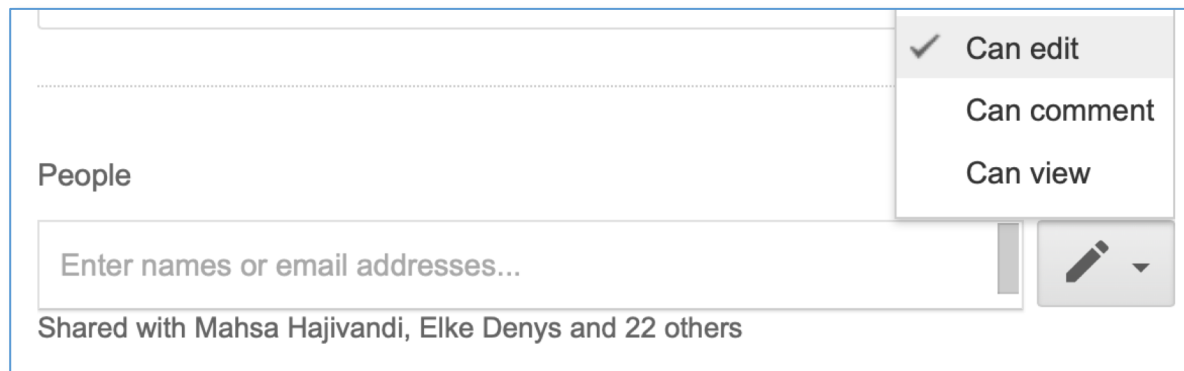
Share with others Get shareable link 

Link sharing on [Learn more](#)

Anyone with the link **can edit** ▼ Copy link

https://docs.google.com/presentation/d/1GkVC3KcQJpSoMlhF9LT-A_RJ45RD5mW

-Dedicated access list



Can edit
Can comment
Can view

People

Enter names or email addresses...

Shared with Mahsa Hajivandi, Elke Denys and 22 others

When would you
use what?

Sharing

Pros & Cons

Shareable link

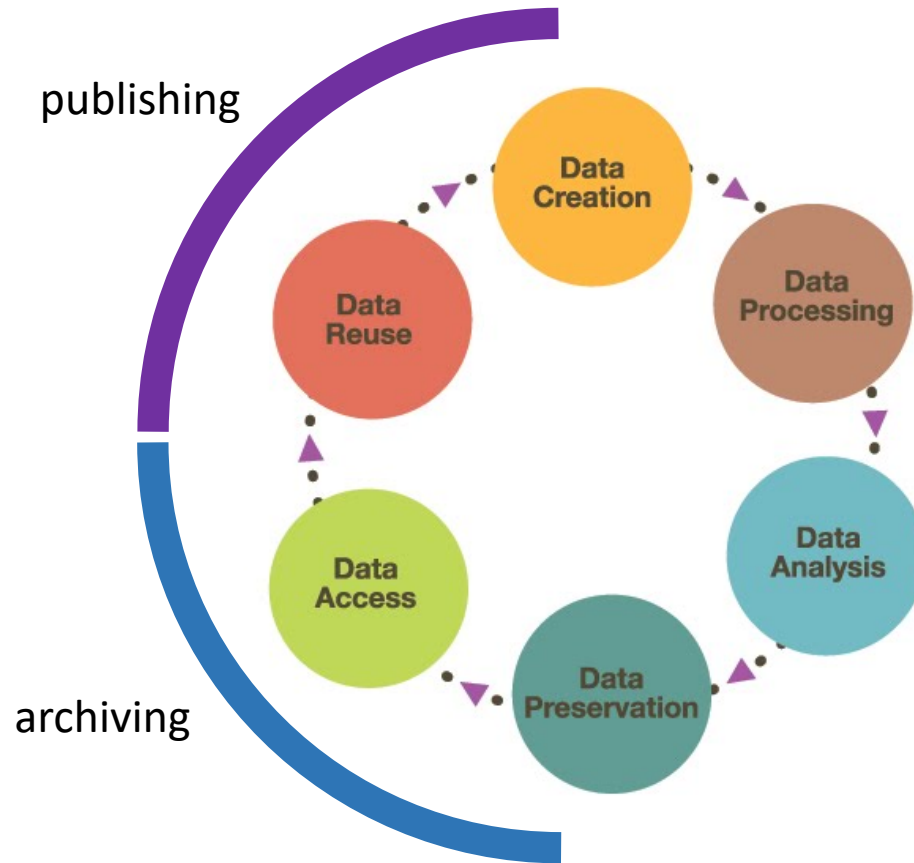
- + Anyone can access the data
- + No need for extra account for collaborators
- Link offers only one granularity of access (read, edit, comment ...)
- Link can be spread further without authors knowledge
- Revoking link revokes access to all collaborators

Dedicated access list

- + Access level can be set per collaborator
- + Access can be revoked for group or individual
- + Data cannot be shared without consent by author
- Every collaborator needs have an account on that system

Terminology....

Sharing versus archiving versus publishing...



Publishing and Archiving

Essential difference:

Publishing: Focus on visibility and accessibility of data and information.

Archiving: Focus on long term preservation and retrievable data and information.

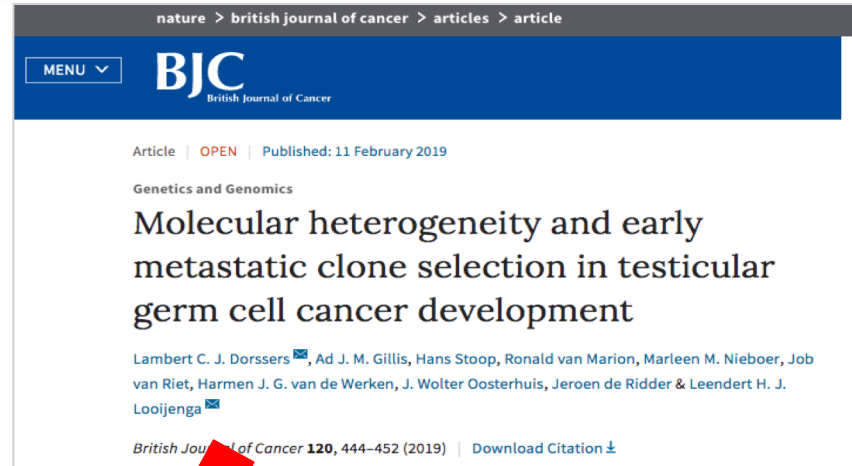
Publishing formats

Data Publishing

- As supplementary data to a publication (no DOI)
- On a project website (no enough metadata)
- Via a standardised data archive (certified and curation workflow) (ideal)
- Via a domain specific data repository (self publishing, no curation workflow) (possible)
- Data paper
- Combinations

Data publishing examples

As **supplementary** data to a publication (no DOI)





nature > british journal of cancer > articles > article


MENU **BJC**
British Journal of Cancer

Article | OPEN | Published: 11 February 2019

Genetics and Genomics

Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development

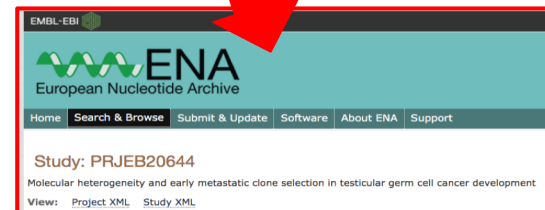
Lambert C. J. Dorssers , Ad J. M. Gillis, Hans Stoop, Ronald van Marion, Marleen M. Nieboer, Job van Riet, Harmen J. G. van de Werken, J. Wolter Oosterhuis, Jeroen de Ridder & Leendert H. J. Looijenga 


British Journal of Cancer **120**, 444–452 (2019) | Download Citation 

Data availability

WGS, ENA PRJEB20644, accession numbers ERX2100523–530; targeted sequencing, ENA PRJEB20644, accession numbers: ERX2019898–958; RNAseq, ArrayExpress, accession number: E-MTAB-5746; DNA methylation, GEO (GSE58538, GSM1413103–GSM1413106) or ArrayExpress (E-MTAB-5842, sample T6107-YSTmeta).

Paper: <https://www.nature.com/articles/s41416-019-0381-1>
Data: <https://www.ebi.ac.uk/ena/data/view/PRJEB20644>



EMBL-EBI 

ENA

European Nucleotide Archive

Home Search & Browse Submit & Update Software About ENA Support

Study: PRJEB20644

Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development

View: [Project XML](#) [Study XML](#)

Data publishing examples

On a **project website**
(not enough metadata)

Dr. Javier Alonso-Mora, Assistant Professor

Autonomous Multi-Robots Lab. Delft University of Technology

[Home](#) | [Research](#) | [Team](#) | [Publications](#) | [Teaching](#) | [Videos and Media](#) | [Bio & Contact](#) |

A complete list of my publications is available in my [Google Scholar](#) profile.

Journals

(J18) H. Zhu and J. Alonso-Mora, "Chance-constrained Collision Avoidance for MAVs in dynamic environments", in IEEE Robotics and Automation Letters (RA-L), Jan. 2019. [\[PDF\]](#) [\[video\]](#)

← → ↻ ⓘ Not Secure | www.alonsomora.com/docs/19-zhu-RAL.pdf

IEEE ROBOTICS AND AUTOMATION LETTERS. PREPRINT VERSION. ACCEPTED DECEMBER, 2018

Chance-Constrained Collision Avoidance for MAVs in Dynamic Environments

Hai Zhu and Javier Alonso-Mora

Abstract—Safe autonomous navigation of micro air vehicles in cluttered dynamic environments is challenging due to the uncertainties arising from robot localization, sensing and motion disturbances. This paper presents a probabilistic collision avoidance method for navigation among other robots and moving obstacles, such as humans. The approach explicitly considers the collision probability between each robot and obstacle and formulates a chance constrained nonlinear model predictive control problem (CCNMPC). A tight bound for approximation of collision probability is developed which makes the CCNMPC formulation tractable and solvable in real time. For multi-robot coordination we describe three approaches, one distributed without communication (constant velocity assumption), one distributed with communication (of previous plans) and one centralized (sequential planning). We evaluate the proposed method in experiments with two quadrotors sharing the space with two humans and verify the multi-robot coordination strategy in simulation with up to sixteen quadrotors.

Index Terms—Path Planning for Multiple Mobile Robots or Agents, Collision Avoidance, Motion and Path Planning.

only the sensed velocity and position of neighboring robots are used, a distributed approach where previous plans of other robots are communicated, and a centralized approach for multi-robot coordination where a sequential planning scheme is employed.

(a) Snapshot from experiment

(b) Schematic of quadrotors, human, plan and uncertainties

Fig. 1: Probabilistic collision avoidance among obstacles.

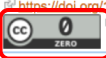
Website: <http://www.alonsomora.com/publications.html>
PDF: <http://www.alonsomora.com/docs/19-zhu-RAL.pdf>

Data publishing examples

Via a standardised
data archive
(certified and
curation workflow)

Dataset | **Genotyping Data Hyperrecombinant offspring VIGS** Link to <https://doi.org/10.4121/uuid:33c2cd57-c041-4b2c-b15b-736b2451e7e1> How to cite this dataset

▼ go to DATA section ▼

title	7	Genotyping Data Hyperrecombinant offspring VIGS
creator	7	orcid Calvo-Baltanás, V. (Vanessa)
creator	7	orcid Wijnker, E. (Erik)
contributor	7	Laboratory of Genetics, Wageningen University & Research
date accepted	7	2019-01-24
date created	7	2017 through 2018
date published	7	2019-01-24
description	7	Offspring generated from the crosses of F1 lrxcol plants in which RECQ4 and/or FIGL1 were presumably knocked-down. Expected lines were expected to be hyperrecombinant but they display the same recombination events as compared to a wild-type population. The genotyping markers used can be seen in the first two rows while the first two columns display the total number of lines used. The markers in blue (B) correspond to homozygous Ler alleles while the green ones (H) correspond to the presence of a Col-Ler alleles.
funder	7	COMREC [Marie Curie]
keyword	7	genetics ◊ genotyping ◊ Increase of recombination VIGS ◊ recombination ◊ virus-induced gene silencing (VIGS)
language	7	en
publisher	7	4TU.Centre for Research Data
subject	7	0604 - Genetics
▲ in collection	7	Datasets of dissertations
related publication	7	https://doi.org/10.18174/467275
licence	7	 CC BY-NC-ND [more info...]

DATA

[Thesis_Genotyping increase_VCB-.xlsx](#) (application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) | MD5: 6b74ebb52642ea7ef558a6b0a6a82ec2
size: 559549 B (546 KiB)

Dataset: <https://data.4tu.nl/repository/uuid:33c2cd57-c041-4b2c-b15b-736b2451e7e1>

Data publishing examples

April 20, 2018

Colour 0.3.13

Mansencal, Thomas; Mauderer, Michael; Parsons, Michael; Canavan, Luke; Cooper, Sean; Shaw, Nick; Wheatley, Kevin; Crowson, Katherine; Lev, Ofek; Leinweber, Katrin; Vandenberg, Jean D.; Sharma, Shriramana

Colour Science for Python

Colour is a Python colour science package implementing a comprehensive number of colour theory transformations and algorithms.

It is open source and freely available under the [New BSD License](#) terms.

Software Open Access

1,860

views

67

downloads

[See more details...](#)

Versions

Version 14	10.5281/zenodo.2647615	Apr 20, 2018
Version 13	10.5281/zenodo.2604314	Mar 19, 2018
Version 12	10.5281/zenodo.1175177	Feb 18, 2018
Version 11	10.5281/zenodo.821825	Jul 12, 2017
Version 10	10.5281/zenodo.376790	Mar 12, 2017

[View all 14 versions](#)

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.605791](#). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Publication date:

April 20, 2018

DOI:

DOI [10.5281/zenodo.2647615](#)

Keyword(s):

API Biochemistry Blackbody Characterisation Chromatic Adaptation Colorimetry Colour Colour Appearance Model Colour Difference Colour Matching Functions Colour Model Colour Notation System Colour Quality Colour Rendition Chart Colour Science Correlated Colour Temperature Illuminants Lightness Luminance Luminous Efficiency Function Open Source Optical Phenomenon Photometry Planckian Radiator Python Reflectance Recovery Spectrum Tristimulus Values Whiteness

Grants:

European Commission:

- DEEVIEW - GAZE-BASED PERCEPTUAL AUGMENTATION (303780)

Related identifiers:

Cites:

<https://github.com/colour-science/colour/blob/develop/BIBLIOGRAPHY.rst>

Communities:

European Commission Funded Research (OpenAIRE)
Zenodo


License (for files):

[BSD 3-Clause "New" or "Revised" License](#)

Dataset: <https://zenodo.org/record/2647615#.XNlHttMzZNO>

Data publishing examples

Data Paper E.g. Data in Brief by ScienceDirect (Elsevier)



Journal
Data in Brief

Volume 7, June 2016, Pages 1451-1454

Data Article

Hand measurement data from human babies at birth, from 26 to 41 weeks estimated gestational age

Emmanuelle Honoré ^{a, b, d}, Thameur Rakza ^c, Philippe Deruelle ^c

Show more

<https://doi.org/10.1016/j.dib.2016.03.089> Get rights and content
Under a Creative Commons license open access

Refers to Emmanuelle Honoré, Thameur Rakza, Brigitte Senut, Philippe Deruelle, Emmanuelle Pouydebat
First identification of non-human stencil hands at Wadi Sūra II (Egypt): A morphometric study f...
Journal of Archaeological Science: Reports, Volume 6, April 2016, Pages 242-247
Download PDF

Specifications table

Subject area	Biology
More specific subject area	Biometry
Type of data	Table
How data was acquired	Medical caliper
Data format	Raw
Experimental factors	Measurements were taken directly on hands with a medical caliper.
Experimental features	Measurements were taken in the first week of life.
Data source location	Neonatology Unit, CHRU Jeanne de Flandre (University Hospital), Lille, France
Data accessibility	Data is with this article

1. Data

Two series of measurements taken at birth on the hands of human newborns are displayed: a series from babies born pre-term, from 26 to 36 weeks EGA (Estimated Gestational Age), and a series from babies born at term, from 37 to 41 weeks EGA. Data was collected in the Neonatal Unit of the CHRU Jeanne de Flandre (University Hospital) in Lille, France, from January until May 2014. Seven measurement criteria were selected, concerning either lengths, widths or ray of the hand, the palm and the digits. They are recorded with the EGA, the sex and the weight of the individuals, regardless of the side – right hand or left hand (Tables 1 and 2).

Table 1. Measurement series on the hand of 25 human babies born pre-term (26–36 EGA), in mm.

NAME	Surname	Estimated gestational age	Weight	Sex	W ₁	W ₂	R ₁	L _m	L _p	L _h	W _h
RAH	OUN	34	1580	M	6	7.67	24.97	23.39	27.85	51.24	26.93
RAH	KAI	34	2125	M	7.85	7.64	32.01	25.13	31.64	56.77	28.73
KER	YOU	31	1100	M	5.82	6.23	25.8	20.96	29.04	50	23.57

Data in Brief:
<https://www.journals.elsevier.com/data-in-brief>
Data Paper:
<https://www.sciencedirect.com/science/article/pii/S2352340916301974>

Data publishing examples

Combined publishing The GBIF example

The screenshot shows the nlbif IPT interface. At the top, it says "Integrated Publishing Toolkit (IPT)" and "Logged in as c.h.j.aof@stowa.nl". The main heading is "Dutch Foundation for Applied Water Research (STOWA) - Limnodata Neerlandica". Below this, there is a summary of the dataset, including keywords like "Limnofauna, aquatic; riparian; plants; phytoplankton; diatoms; zooplankton; macrofauna; fish; Netherlands; monitoring". It also lists metadata such as "Resource Language: English", "Last Publication: Version 3 from Nov 5, 2012", and "Darewin Core Archive: download (89 MB) 2,942,518 records". There are sections for "External Links", "Resource Contact" (Bas van der Wal), and "Resource Creator" (Roel Kroonen).

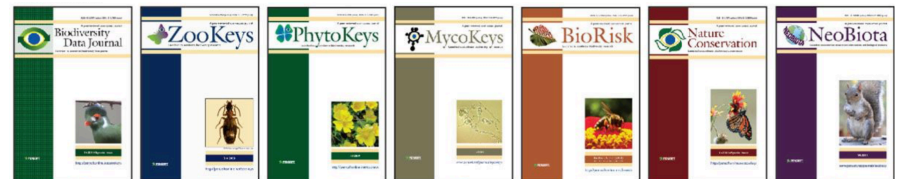
Integrated
Publishing
Toolkit

The screenshot shows the GBIF portal page for the dataset. The header includes "GBIFORG" and "Dutch Foundation for Applied Water R...". A green banner displays "2,942,518 Occurrences" and a "View occurrences" button. The "Summary" section provides details: "FULL TITLE: Dutch Foundation for Applied Water Research (STOWA) - Limnodata Neerlandica", "DESCRIPTION: The Limnodata Neerlandica data set contains the data of more than 30 years systematic and project based sampling of Dutch, mainly freshwater, waterbodies...", "PURPOSE: Water management and research.", "TEMPORAL COVERAGES: Date range: 1-jan-1980 - 31-dec-2010", and "LANGUAGE OF DATA: English". It also lists administrative contact, metadata author, and originator.

GBIF portal www.gbif.org

Every dataset & search a DOI

Pensoft data papers



With the same toolkit:

- Data online
- Submitting data papers

Data publishing examples

Which way of data publication is your preferred one?

1. Supplementary material to paper
2. Own webserver
3. Certified data repository with review
4. Self-publishing through data repository
5. Peer-reviewed data publication (Data in brief, F1000, ...)

Which data publication would you trust?

With which method would you reach the biggest audience?

Terminology....

What is a repository?

Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis.



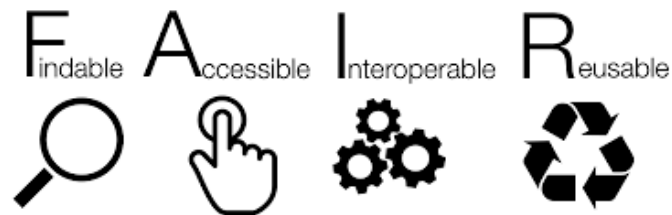
<https://dictionary.casrai.org/>

Terminology....

What is a repository?

Repositories **preserve**, **manage**, and provide **access** to many types of digital materials in a **variety of formats**. Materials in online repositories are **curated** to enable search, discovery, and reuse. There must be **sufficient control** for the digital material to be **authentic**, reliable, accessible and usable on a **continuing** basis.

Good repository means FAIR data



<https://dictionary.casrai.org/>

Publishing platforms

Overview

Many repositories

- Multi purpose repositories like B2SHARE, public version of Figshare and Zenodo
 - Simple publishing workflow, depositor is responsible for content, quality and metadata of the published data
- Curated repositories like DANS EASY, 4TU Data centre repository, many institutional repositories
 - Depositor sends data over, then data is checked for quality to some extent
- Community-specific repositories like EBI EGA, NCBI GEO, ...
 - Expect very specific data formats
 - Have an extended data quality checking pipeline (people try to reproduce the data)

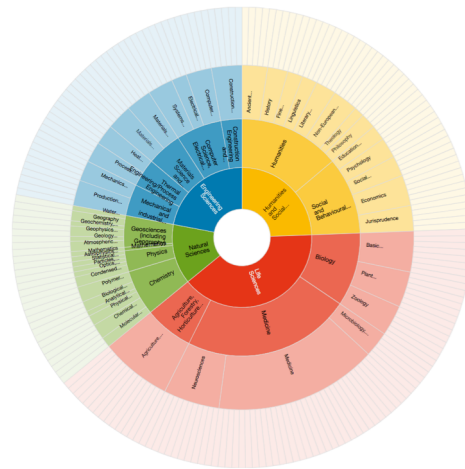
Publishing platforms

Many Repositories - re3data.org

Re3data.org

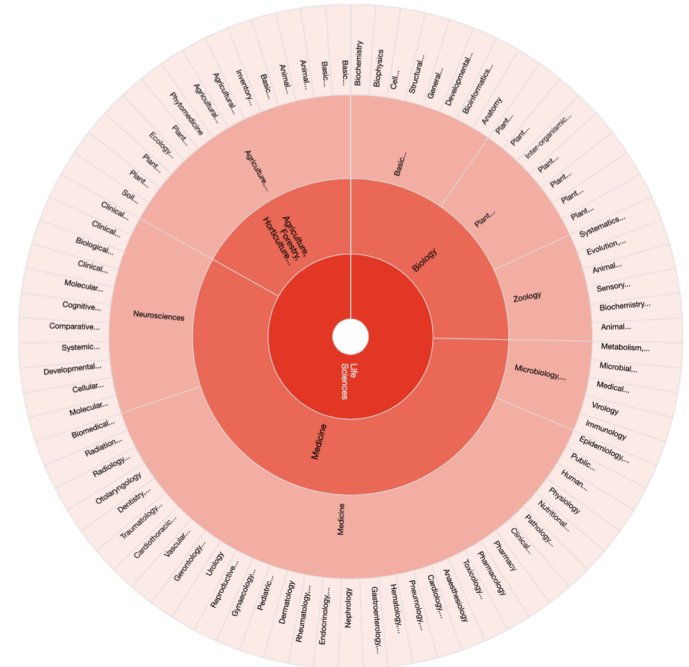
Over 2300 international repositories for a wide spread of domains

Plentiful filters to find appropriate repositories



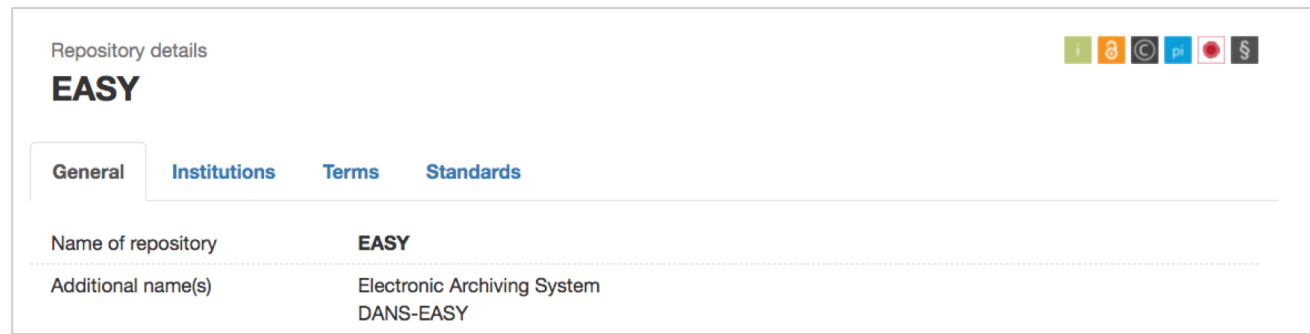
All Categories

Life Science
Repositories
Categories



Publishing platforms

Many Repositories - re3data.org



The screenshot shows the 'Repository details' page for 'EASY' on re3data.org. The page has a header with the repository name 'EASY' and a row of social media icons. Below the header are four tabs: 'General', 'Institutions', 'Terms', and 'Standards'. The 'General' tab is selected, showing the following information:

Name of repository	EASY
Additional name(s)	Electronic Archiving System DANS-EASY

Provides overview of

- general information about the individual features,
- institutional characteristics,
- terms of use and
- applied standards

Publishing platforms re3data.org

Example: CERN ZENODO



Type:

General Repository for open source, open access, open data

Publishing Policy:

Once it's published the record cannot be deleted; metadata can be adjusted at any time

Advantage:

Quick and independent publishing
Published with one button click
Standardised repository that provided DOI

Disadvantage:

Typos and writing errors are not double checked
Errors and issues on file-level are not detected

A screenshot of the Zenodo website interface. At the top, a black banner contains the text "The research data repository is neither certified nor supports a repository standard." Below this is a navigation bar with the Zenodo logo, a search box, and links for "Upload" and "Communities". A "File Type" list is visible on the left, showing various file formats and their counts. On the right, a section titled "Zenodo in a nutshell" lists key features: Research. Shared., Citeable. Discoverable., Communities, Funding, Flexible licensing, and Safe. A link to "features" is provided at the bottom of this section.

The research data repository is neither certified nor supports a repository standard.

re3data.org

zenodo Search Upload Communities

File Type

- Pdf (715388)
- Png (151656)
- Jpg (85043)
- Zip (45613)
- Hdf5 (15048)
- Xml (9157)
- Docx (6775)
- Txt (4163)
- Json (4046)
- Gz (3536)

Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citeable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- **Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** — because not everything is under Creative Commons.
- **Safe** — your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

Read more about Zenodo and its [features](#).

Publishing platforms re3data.org

Example: EBI EGA



Type:

Data Archive for sequence and genotype studies and data

Publishing Policy:

Once it's published the data-set cannot be deleted; Data Access Committee (DAC) controls access

Advantage:

Dedicated domain specific archive for Europe
Secure access control via DAC

Disadvantage:

No persistent link
Elaborate upload process

The research data repository is neither certified nor supports a repository standard.

The research data repository does not use a persistent identifier system.

re3data.org

Dataset ID ^	Description v	Technology v	Samples v
EGAD000000000001	WTCCC1 project samples from 1958 British Birth Cohort	Affymetrix 500K,unknown	1504
EGAD000000000002	WTCCC1 project samples from UK National Blood Service	Affymetrix 500K,unknown	1500
EGAD000000000003	WTCCC1 project Bipolar Disorder (BD) samples	Affymetrix 500K	1998
EGAD000000000004	WTCCC1 project Coronary Artery Disease (CAD) samples	Affymetrix 500K	1998
EGAD000000000005	WTCCC1 project Inflammatory Bowel Disease (IBD) samples	Affymetrix 500K	2005
EGAD000000000006	WTCCC1 project Hypertension (HT) samples	Affymetrix 500K	2001
EGAD000000000007	WTCCC1 project Rheumatoid arthritis (RA) samples	Affymetrix 500K	1999
EGAD000000000008	WTCCC1 project Type 1 Diabetes (T1D) samples	Affymetrix 500K	2000
EGAD000000000009	WTCCC1 project Type 2 Diabetes (T2D) samples	Affymetrix 500K	1999
EGAD000000000010	WTCCC1 project Ankylosing Spondylitis (AS) samples	Illumina 15K	957

Moving to a demo session on Dataverse.....

By Christine Staiger

Cees.Hof@dans.knaw.nl

 @CeesH_DANS