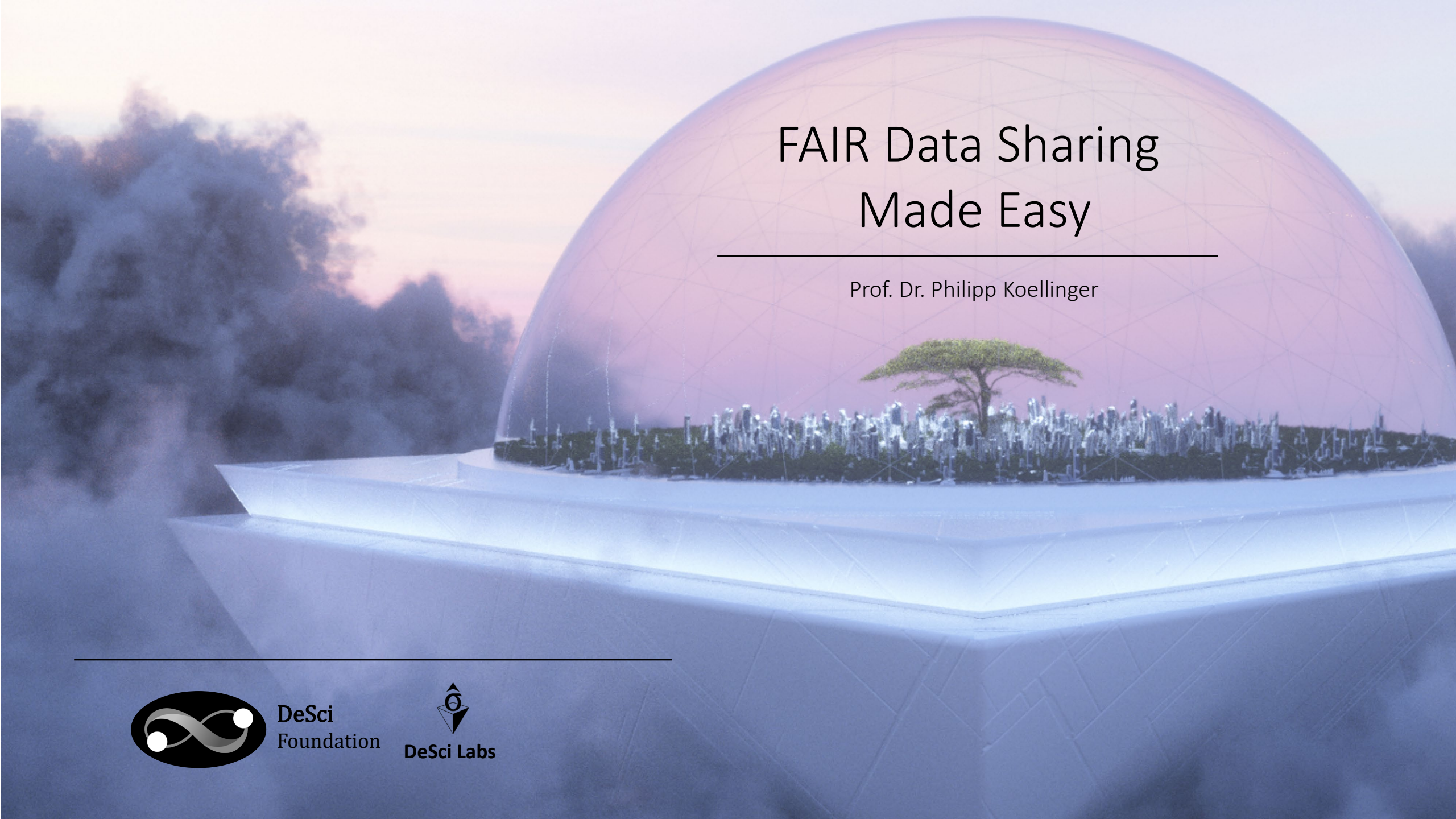


FAIR Data Sharing Made Easy

Prof. Dr. Philipp Koellinger

DeSci Foundation

DeSci Labs

Frustrations of researchers

Rank ↑	Opinion	Score	Actions
1st	Peer-review work is unpaid.	76	
2nd	It's hard to find out if the results of a study are trustworthy (e.g. replicable).	66	
3rd	Many empirical papers do not offer easy access to their underlying data or code.	66	
4th	The peer-review and journal publication process is too slow.	65	
5th	Incentives for independent replication efforts are missing.	60	
6th	Submitting scientific content is a cumbersome process (Why don't submission system simply extract the relevant author information from the manuscript?)	57	
7th	Best practices such as sharing data and code lack formal recognition.	55	
8th	Scientists do not get royalty payments for influential work they publish.	52	
9th	The quality of peer-review reports is often too low.	51	
10th	Peer review happens too late in the research process.	50	
11th	I fear sharing my research too early because my work might get stolen.	44	
12th	Peer-review work lacks formal recognition.	43	
13th	It's difficult to find quality research outputs (papers, data, code) that are relevant to my own work.	40	
14th	Version control of documents and data files is cumbersome and imperfect.	39	
15th	Broken links to scientific content.	35	
16th	Most peer reviews are inaccessible and lost.	33	
17th	It's difficult to find good referees.	24	

DeSci Labs survey results from Jan 2024
N = 94 active researchers, mostly early-career

DeSci Labs

@DesciLabs
@DesciFoundation
@PKoellinger

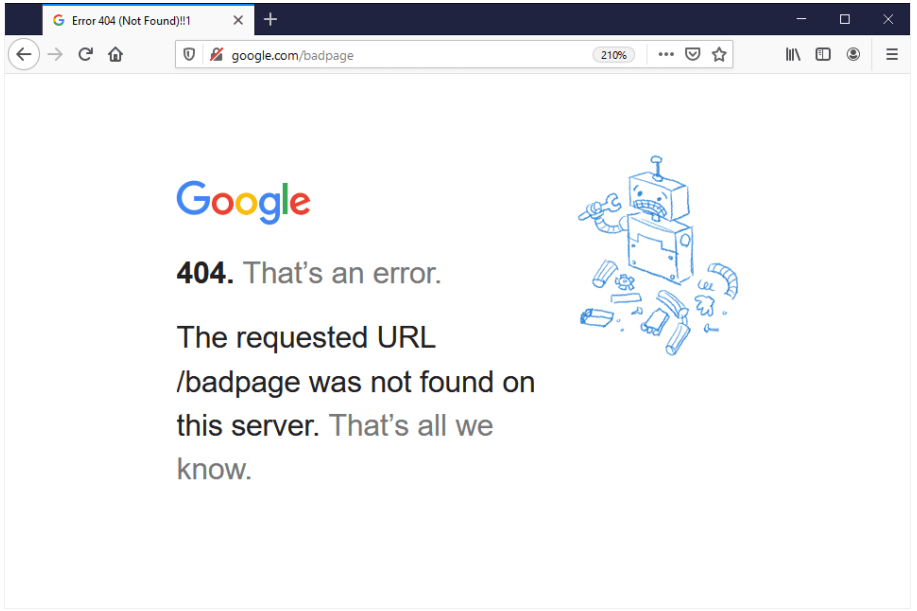
DeSci Foundation

Data sharing pains

- “Data available upon request”
  - No response from authors
  - Authors can’t find their data anymore or have lost access
  - Proprietary file formats (e.g. SPSS, Stata, Eviews)
  - Low-quality meta-data
    - Unclear variable names or labels
  - Missing variables
  - Not the correct version of the data
  - Data are protected & without realistic access path

Data sharing pains

- “Data available upon request”
  - No response from authors
  - Authors can’t find their data anymore or have lost access
  - Proprietary file formats (e.g. SPSS, Stata, Eviews)
  - Low-quality meta-data
    - Unclear variable names or labels
  - Missing variables
  - Not the correct version of the data
  - Data are protected & without realistic access path
  - Link rot (404 error) and content drift
  - Months later…



Data sharing pains

- Big data pains
  - Storage costs
  - Egress costs
  - Moving big data takes long
  - “Data has gravity"

Omer Weissbrod @oweissb · Feb 24, 2023

#statisticalgenetics Twitter we need your help to support open science! We can't afford releasing scientific data 😟

In 2020 we (Alkes Price and me) publicly released UK Biobank LD matrices (3TB of data). In recent months the costs for user downloads approached $100K...
1/2

26 42 91 40K

Omer Weissbrod
@oweissb

Alkes's group can't sustainably afford such costs...

The data is currently on the @broadinstitute cloud. Where we can store the data at reasonable costs? Unless we find a solution we'll have to block access to the data...

2/2

3:41 PM · Feb 24, 2023 · **3,825** Views

But why should I share my data?

- Increase the impact of your work
  - Articles with posted data receive *much* more citations
- Be an open-science leader
- Funding agency & publication requirements
- It's the right thing to do
  - Publicly funded resources should remain public goods
- "As open as possible, as closed as necessary"
  - Access paths to securely stored sensitive data

Citation advantages of papers with data

- Increase the impact of your work
  - Articles with posted data receive *much* more citations
- Example: PLOS and BMC
  - 9% “available upon request or similar”
  - 6% “in paper or SI”
  - 25% “in repository
    - Citation advantage according to regression analysis (Table 6)

PLOS ONE

RESEARCH ARTICLE

The citation advantage of linking publications to research data

Giovanni Colavizza[1,2], Iain Hrynaszkiewicz[3,4], Isla Staden[1,5], Kirstie Whitaker[1,6], Barbara McGillivray[1,6]*

1 The Alan Turing Institute, London, United Kingdom, 2 University of Amsterdam, Amsterdam, Netherlands, 3 Springer Nature, London, United Kingdom, 4 Public Library of Science, Cambridge, United Kingdom, 5 Queen Mary University, London, United Kingdom, 6 University of Cambridge, Cambridge, United Kingdom

* bmcgillivray@turing.ac.uk

Abstract

Efforts to make research results open and reproducible are increasingly reflected by journal policies encouraging or mandating authors to provide data availability statements. As a consequence of this, there has been a strong uptake of data availability statements in recent literature. Nevertheless, it is still unclear what proportion of these statements actually contain well-formed links to data, for example via a URL or permanent identifier, and if there is an added value in providing such links. We consider 531,889 journal articles published by PLOS and BMC, develop an automatic system for labelling their data availability statements according to four categories based on their content and the type of data availability they display, and finally analyze the citation advantage of different statement categories via regression. We find that, following mandated publisher policies, data availability statements become very common. In 2018 93.7% of 21,793 PLOS articles and 88.2% of 31,956 BMC articles had data availability statements. Data availability statements containing a link to data in a repository—rather than being available on request or included as supporting information files—are a fraction of the total. In 2017 and 2018, 20.8% of PLOS publications and 12.2% of BMC publications provided DAS containing a link to data in a repository. We also find an association between articles that include statements that link to data in a repository and up to 25.36% (± 1.07%) higher citation impact on average, using a citation prediction model. We discuss the potential implications of these results for authors (researchers) and journal publishers who make the effort of sharing their data in repositories. All our data and code are made available in order to reproduce and extend our results.

Introduction

More research funding agencies, institutions, journals and publishers are introducing policies that encourage or require the sharing of research data that support publications. Research data policies in general are intended to improve the reproducibility and quality of published research, to increase the benefits to society of conducting research by promoting its reuse, and

OPEN ACCESS

Citation: Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data. PLoS ONE 15(4): e0230416. https://doi.org/10.1371/journal.pone.0230416

Editor: Jelte M. Wicherts, Tilburg University, NETHERLANDS

Received: July 5, 2019

Accepted: February 28, 2020

Published: April 22, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0230416

Data Availability Statement: Code and data can be found at: https://doi.org/10.5281/zenodo.3470062.

Funding: This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and by Macmillan Education Ltd, part

Citation advantages of papers with data

- Increase the impact of your work
  - Articles with posted data receive *much* more citations
- Example: American Economic Review & American Journal of Political Science
  - 100% more citations
    - Citation advantage according to 2SLS regression
    - Using journal policy change as IV (Table 3)

PLOS ONE

RESEARCH ARTICLE

A study of the impact of data sharing on article citations using journal policies as a natural experiment

Garret Christensen[1]*, Allan Dafoe[2], Edward Miguel[3], Don A. Moore[3], Andrew K. Rose[3]

1 U.S. Census Bureau, Washington, DC, United States of America, 2 University of Oxford, Oxford, England, United Kingdom, 3 University of California, Berkeley, California, United States of America

* garret@berkeley.edu

Abstract

This study estimates the effect of data sharing on the citations of academic articles, using journal policies as a natural experiment. We begin by examining 17 high-impact journals that have adopted the requirement that data from published articles be publicly posted. We match these 17 journals to 13 journals without policy changes and find that empirical articles published just before their change in editorial policy have citation rates with no statistically significant difference from those published shortly after the shift. We then ask whether this null result stems from poor compliance with data sharing policies, and use the data sharing policy changes as instrumental variables to examine more closely two leading journals in economics and political science with relatively strong enforcement of new data policies. We find that articles that make their data available receive 97 additional citations (estimate standard error of 34). We conclude that: a) authors who share data may be rewarded eventually with additional scholarly citations, and b) data-posting policies alone do not increase the impact of articles published in a journal unless those policies are enforced.

Introduction

Verifiability and replicability are fundamental to science. The Royal Society's motto "*nullius in verba*" ("take nobody's word for it") encourages scientists to verify the claims of others. By sharing data, scientists can increase the verifiability and credibility of their claims. Most academic journals and professional societies encourage researchers to share their data, but these are often informal recommendations; until recently, few journals required it.

The ease of posting data on the internet has lowered the cost of data sharing; accordingly, advocates of open science have argued that data posting should be standard practice [1], and a growing number of scientific journals have started requiring that authors publicly post their data. However, this requirement remains more the exception than the rule in many fields, and researchers have not routinely posted their data unless journals require them to do so [2–4].

Researchers give several reasons for their failure to post data. Some highlight costs to the individual, including the effort required, the potential for being scooped, and the risk of being

OPEN ACCESS

Citation: Christensen G, Dafoe A, Miguel E, Moore DA, Rose AK (2019) A study of the impact of data sharing on article citations using journal policies as a natural experiment. PLoS ONE 14(12): e0225883. https://doi.org/10.1371/journal.pone.0225883

Editor: Florian Naudet, University of Rennes 1, FRANCE

Received: June 27, 2019

Accepted: November 14, 2019

Published: December 18, 2019

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0225883

Data Availability Statement: All data files are available on the Open Science Framework: https://osf.io/pxdch/ and https://osf.io/cdt8y/.

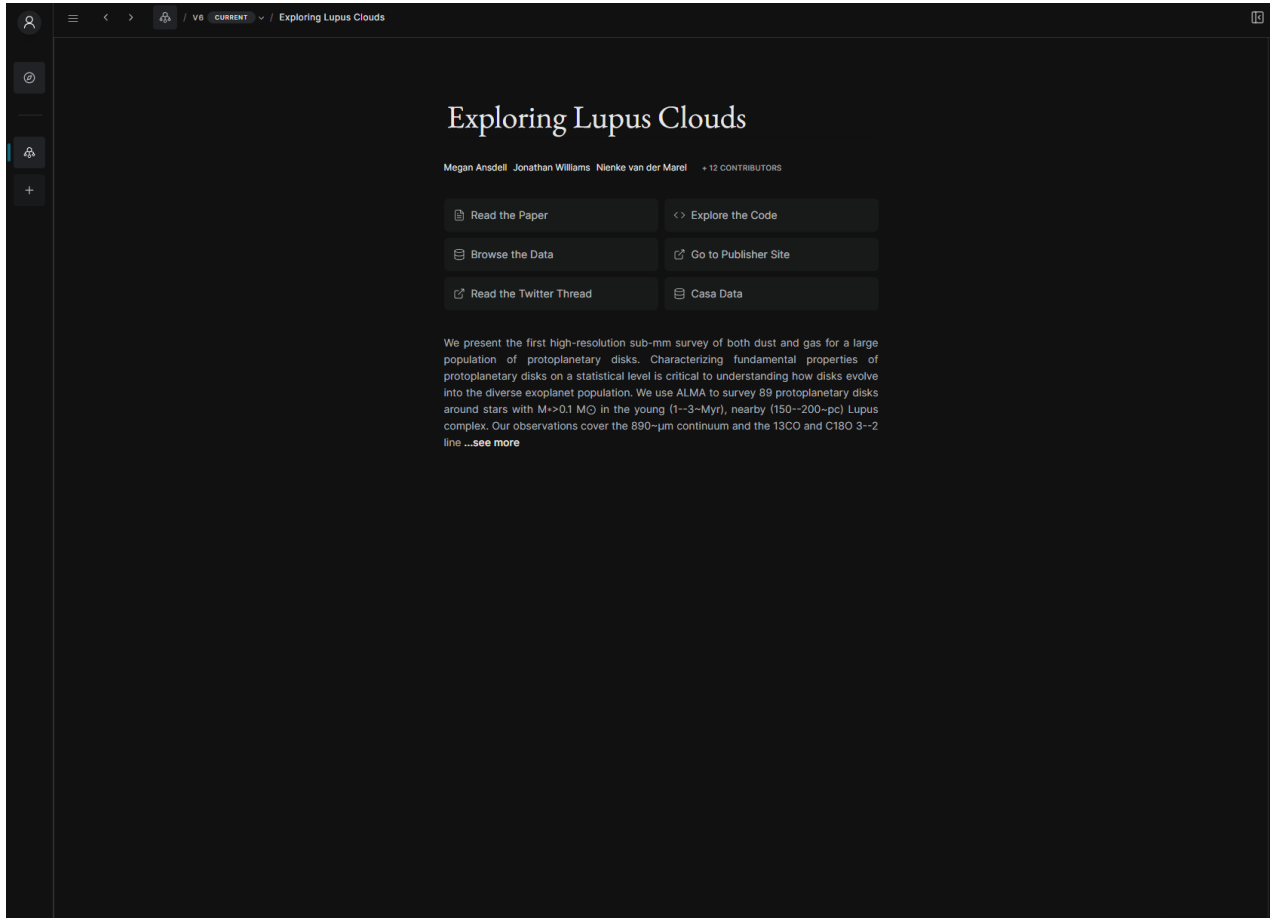
FAIR data

- FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable
  - https://www.gofair.foundation/interpretation
- Requires persistent identifiers for every file
  - File paths & URLs are affected by link rot and content drift
  - DOIs are better
  - Cryptographic fingerprints of files are best
    - Content-addressed data storage based on hash functions (e.g. SHA256)
- Requires high-quality metadata
  - Ideally readable for both humans and machines
  - Meaningful variable names & labels
  - Controlled vocabularies & ontologies

Easy data sharing with DeSci Nodes leads to better data re-use

- Up to 100GB free
- Manuscripts, data, code etc. all in one place
  - Easy data drive
- Versionability
  - Keep track of changes
- Automatic persistent identifiers for each file (dPIDs)
  - No link rot or content drift
- Stored on an open peer-to-peer network (IPFS)
  - It's your choice where you store your data
- Compute-over-data
- Programmatic importing of data & code from Nodes to local compute environments
- FAIR meta-data made easy
- CLI access to build PIDs
- Earn attestations and rewards for your work
  - Show them on your ORCID profile

Demo



DeSci Foundation

@DesciLabs
@DesciFoundation
@PKoellinger

DeSci Labs

Prof. Dr. Philipp Koellinger

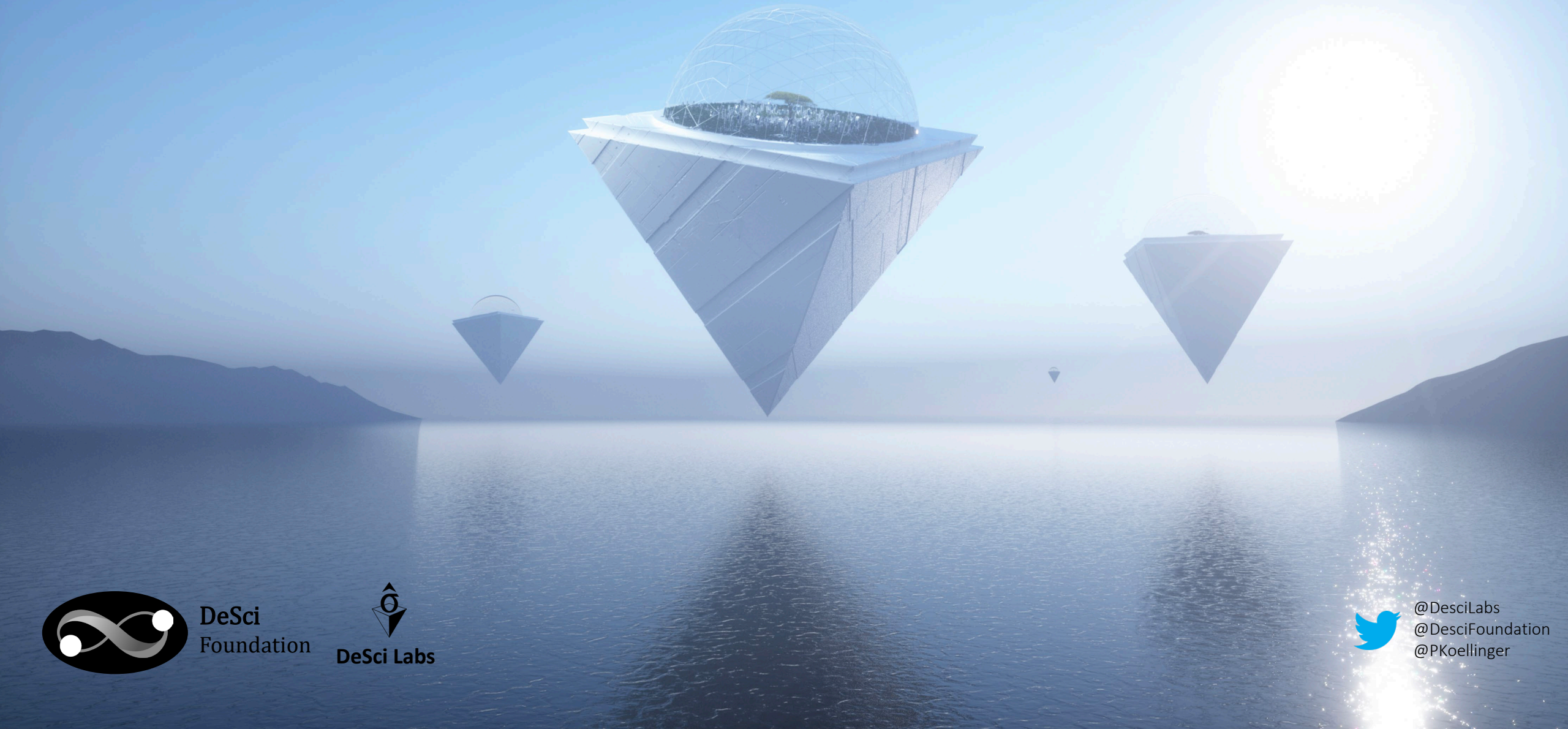
p.d.koellinger@vu.nl

DeSci Nodes

nodes.desci.com

Future of Science Seminar & Podcast

https://descifoundation.org/seminar



DeSci Foundation

DeSci Labs

@DesciLabs
@DesciFoundation
@PKoellinger