

Datas Importantes

Atos acadêmicos no SIGA - Calendário Trimestral	1º Trimestre	2º Trimestre	3º Trimestre	4º Trimestre
Início de atividades	06/07/2020	13/10/2020	01/02/2021	----- X -----
Rematricula de matrícula trancada (destrancamento de matrícula)	Até 27/06/2020	Até 05/10/2020	Até 23/01/2021	----- X -----
Previsão de turmas	Até 19/06/2020	Até 26/09/2020	Até 15/01/2021	----- X -----
Trancamento de matrícula	Até 10/08/2020	Até 17/11/2020	Até 08/03/2021	----- X -----
Pedido de inscrição em disciplinas	De 06/07/2020 a 24/07/2020	De 11/10/2020 a 17/10/2020	De 30/01/2021 a 05/02/2021	----- X -----
Concordância do pedido de inscrição em disciplina	De 27/07/2020 a 30/07/2020	De 18/10/2020 a 24/10/2020	De 06/02/2021 a 12/02/2021	----- X -----
Efetivação do Pedido de Inscrição (Divisão de Ensino – PR2)	31/07/2020	27/10/2020	15/02/2021	----- X -----
Pedido de alteração de inscrição em disciplina – AID	De 02/08/2020 a 08/08/2020	De 28/10/2020 a 31/10/2020	De 16/02/2021 a 19/02/2021	----- X -----
Concordância do pedido de alteração de inscrição em disciplina - AID	De 09/08/2020 a 12/08/2020	De 01/11/2020 a 07/11/2020	De 20/02/2021 a 26/02/2021	----- X -----
Efetivação De Alteração do Pedido de Inscrição (Divisão de Ensino – PR2)	13/08/2020	10/11/2020	01/03/2021	----- X -----
Pedido de trancamento de inscrição em disciplina (desistência de inscrição)	De 14/08/2020 a 19/08/2020	De 11/11/2020 a 14/11/2020	De 02/03/2021 a 05/03/2021	----- X -----
Concordância do pedido de trancamento de inscrição em disciplina	De 20/08/2020 a 31/08/2020	De 15/11/2020 a 28/11/2020	De 06/03/2021 a 19/03/2021	----- X -----
Efetivação do Trancamento do Pedido de Inscrição (Divisão de Ensino – PR2)	24/08/2020	01/12/2020	22/03/2021	----- X -----
Término de atividades	03/10/2020	16/01/2021	24/04/2021	----- X -----
Notas – Pautas de graus e frequência	De 04/10/2020 a 17/10/2020	De 17/01/2021 a 30/01/2021	De 25/04/2021 a 08/05/2021	----- X -----

Entregáveis – Março

4/3 - Aula + Definição dos grupos + PIT 5 min (Apresentação geral em PPT - DataSet + Problema + Objetivo)

11/3 - Aula + Descrição do projeto de DS - Tudo é via GIT!

- Entregar o projeto contendo (V1): Detalhamento e descrição textual da definição do problema a ser explorado pela equipe; descrever tecnicamente o dataset e sua fonte, o que pretendem fazer e o que vão e como extrair (plano dos experimentos). Apresentar os objetivos geral e específico do projeto de DS; apresentar métodos de data cleaning/tratamento de dados que serão usados, apresentar a proposta de modelo de extração de conhecimento e visualização de dados que será adotado.

18/3 - Aula + Refinamento do projeto de DS

- Agregar ao projeto (V2) : Plano do experimento a ser executado (scripts no Colab x Jupyter x IDE), projeto de coleta de metadados da proveniência dos experimentos, projeto para tornar seu experimento reproduzível, adicionar qualquer outro melhoria ou diferencial que julguem necessário

25/3 – Aula + entregar da 1a versão do artigo – Recomenda-se usar o template da LNCS e suas regras de formatação.

- Texto final (15-20) páginas com referências (pode ser em português ou inglês, escolha do grupo)

Entregáveis - Abril

1/4 - Aula

8/4 - Aula + Sorteio ordem de apresentação dos grupos

15/4 - Entregar da 2ª versão do artigo + Apresentação trabalho alunos (1ª. Parte dos grupos)

22/4 - Apresentação - trabalho alunos (2ª. Parte dos grupos)

10/5 (?) - Entrega da versão final do artigo + projeto completo (V3) + scripts com provenance + datasets anotados + resultados de DS + Depósito do notebook reproduzível e executáveis no GIT



PPGI PROGRAMA
DE PÓS-GRADUAÇÃO
EM INFORMÁTICA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Introduction to Data Science

MODULE IV – PART III

FAIR Data Science

Prof Sergio Serra e Jorge Zavaleta

What? The Reproduction Process

Ask an interesting question
& learn reproducibility

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

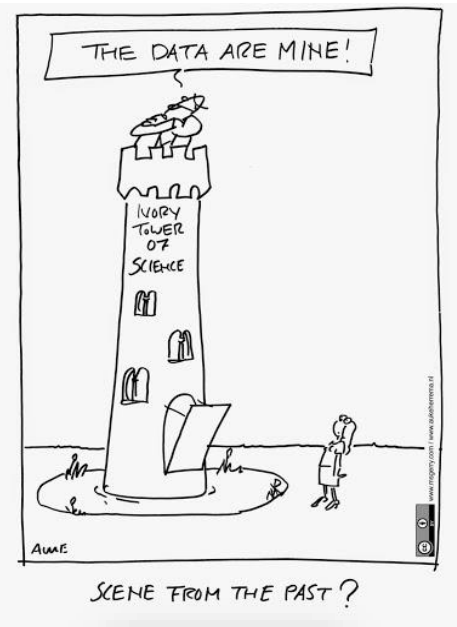
How to make my DS project FAIR?

Module IV

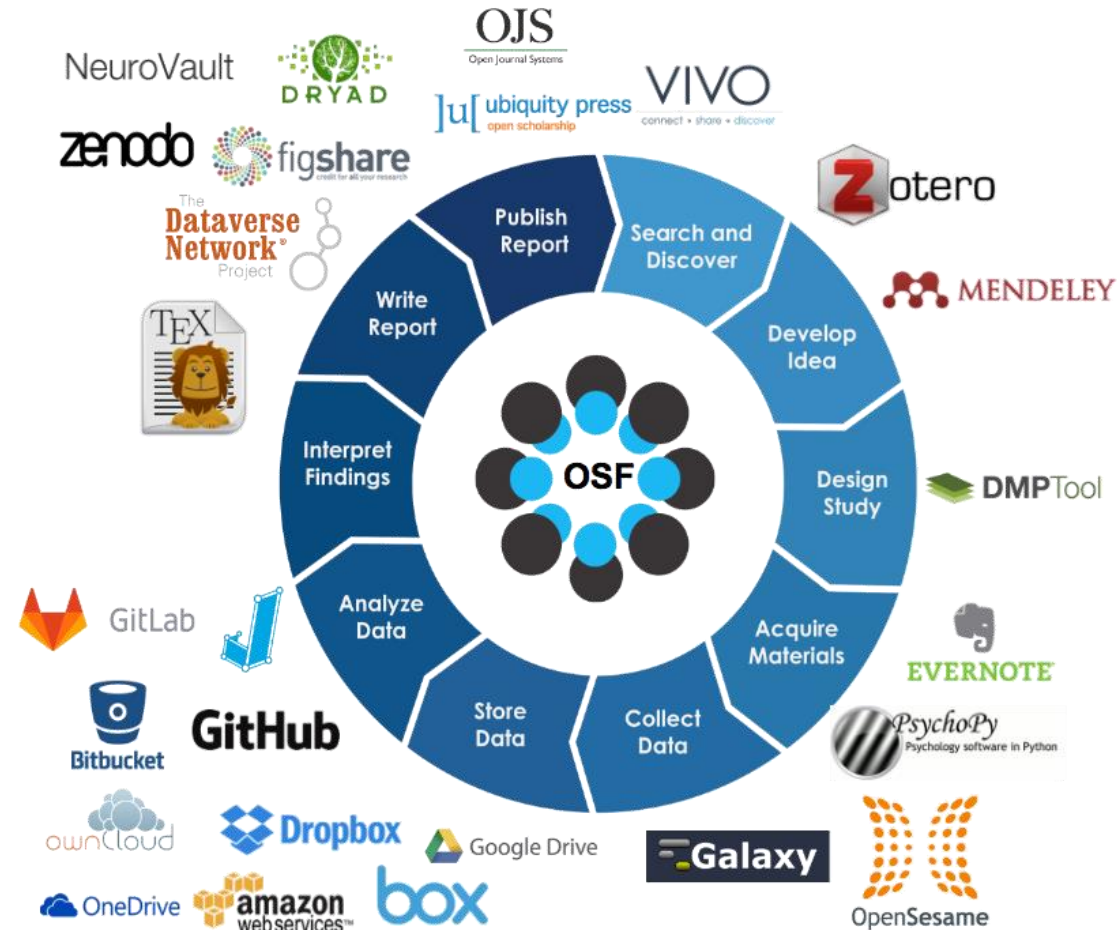
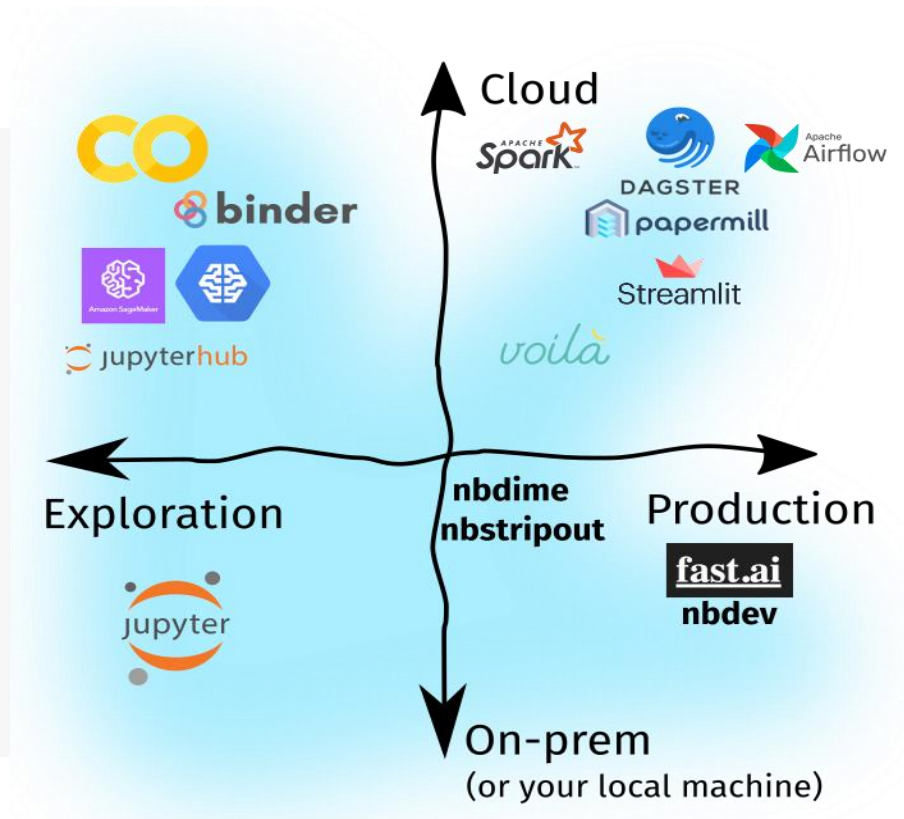
The data science landscape today



The data science landscape today



Be ambitious 🔍



A simplified framework on how to think about the advancements in the data science process for the past years

Data Policies

Data Science

Data Skills

Data Good Practices



- CODATA Data Policy Committee
<http://bit.ly/data-policy-committee>;
- One major policy report per year.
- 20-Year Review of GBIF currently underway.
- New Centre of Excellence in Data for Society being set up at University of Arizona.



- Data Science Journal:
<https://datascience.codata.org/>
- International Data Week and CODATA Conference series.
- Task Groups and Working Groups.

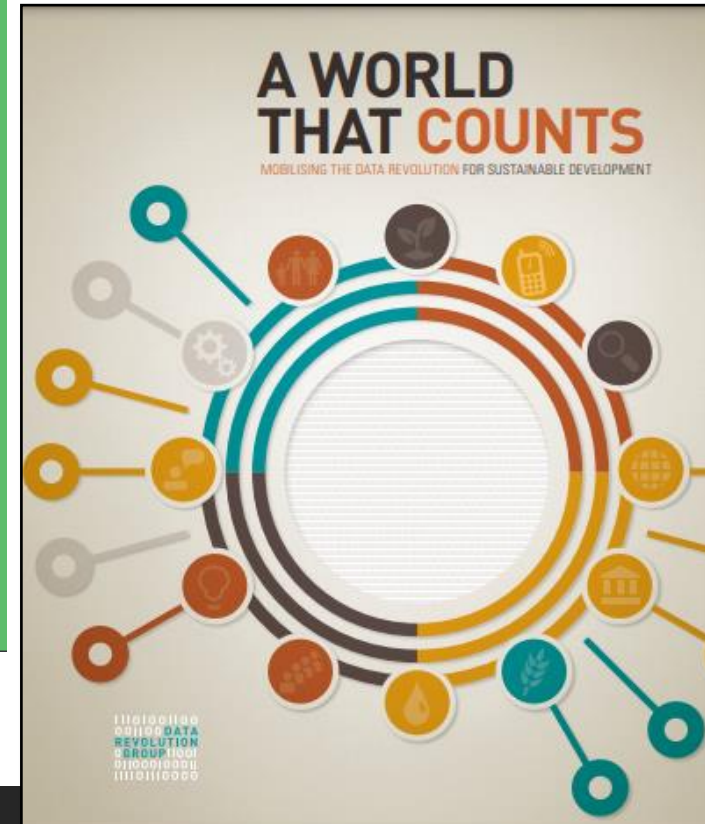
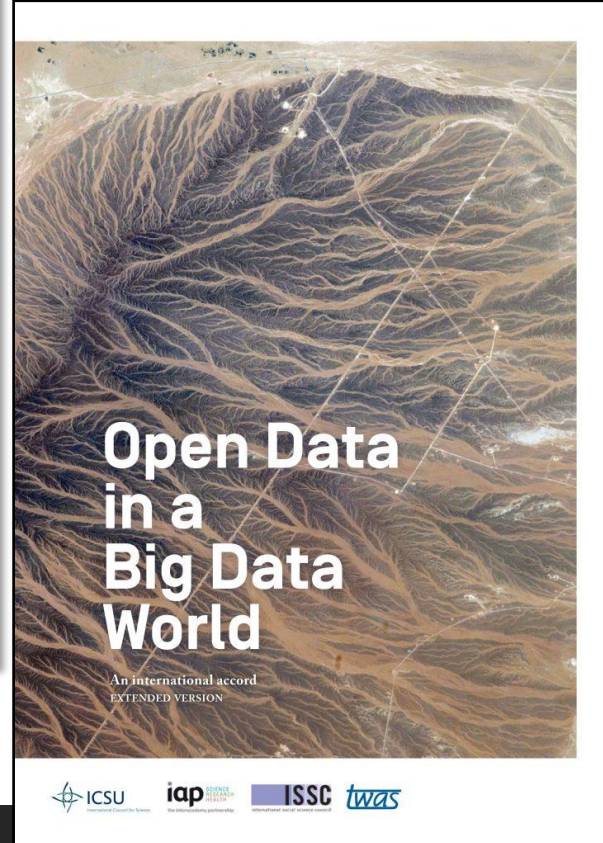
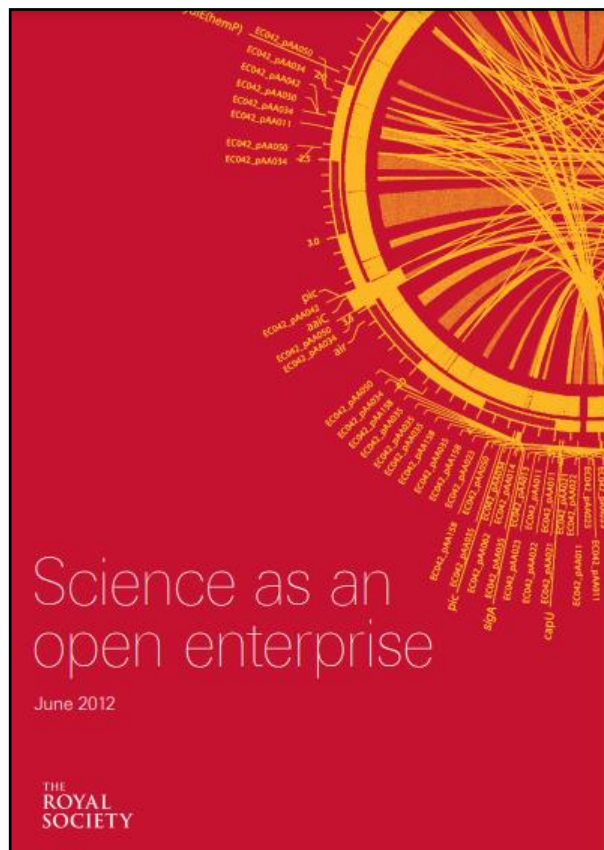


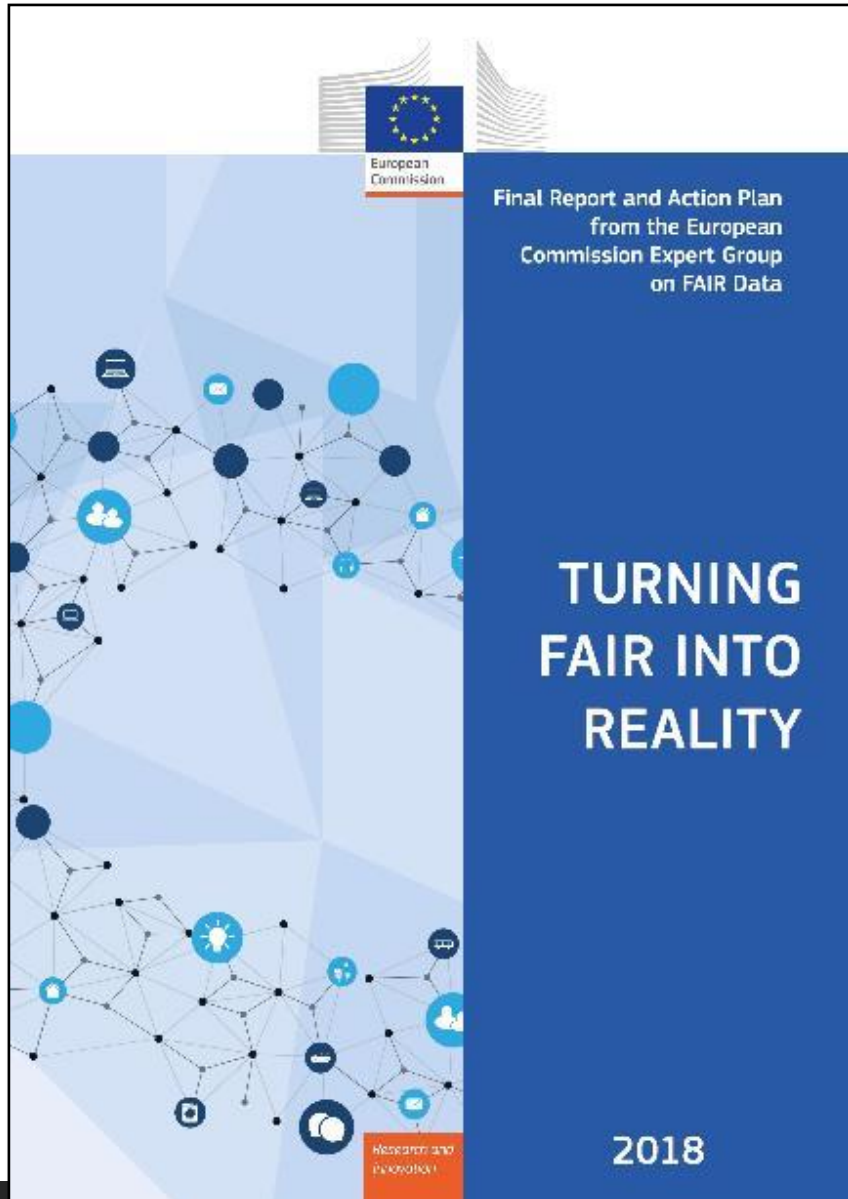
- CODATA-RDA School of Research Data Science.
- CODATA China, PASTD and other training activities.
- #terms4FAIRskills and FAIRsFAIR Competence Centres.



- Regional Open Science Platforms
- Data Interoperability for Multi-Disciplinary Research.
- Survey and recommendation of good practices.

Data Science + Open Science + FAIR Data





FAIR

- **Findable:** sufficiently rich metadata and a unique and persistent identifier, to enable discovery.
- **Accessible:** retrievable by humans and machines through a standard protocol; authentication and authorization where necessary.
 - Allows programmatic access for analysis.
- **Interoperable:** metadata use a 'formal, accessible, shared, and broadly applicable language for knowledge representation'.
 - The descriptions of variables etc follow a shared specification and are commensurable.
- **Reusable:** metadata provide rich and accurate information; clear usage license; detailed provenance.
 - Both humans and their analytical tools know what can be done with the data (license) and can assess its provenance.

FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

(Mons, B., et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, <http://dx.doi.org/10.1038/sdata.2016.18>)



FAIR and Open, but not a Silver Bullet!

- **FAIR \neq Open \neq Data Quality:** FAIR and Open and Quality are distinct concepts but complementary.
- **Drivers for FAIR:**
 - not enough to make Data Open, dump it raw onto the Web
 - important to have a dialogue with research areas in which much data cannot be Open
- FAIR is useful because it applies as much to data that **MUST** be restricted as to data that can be Open
- FAIR does **NOT** detract from Open
- Research data should be as Open as possible, Open by default.



High level observations

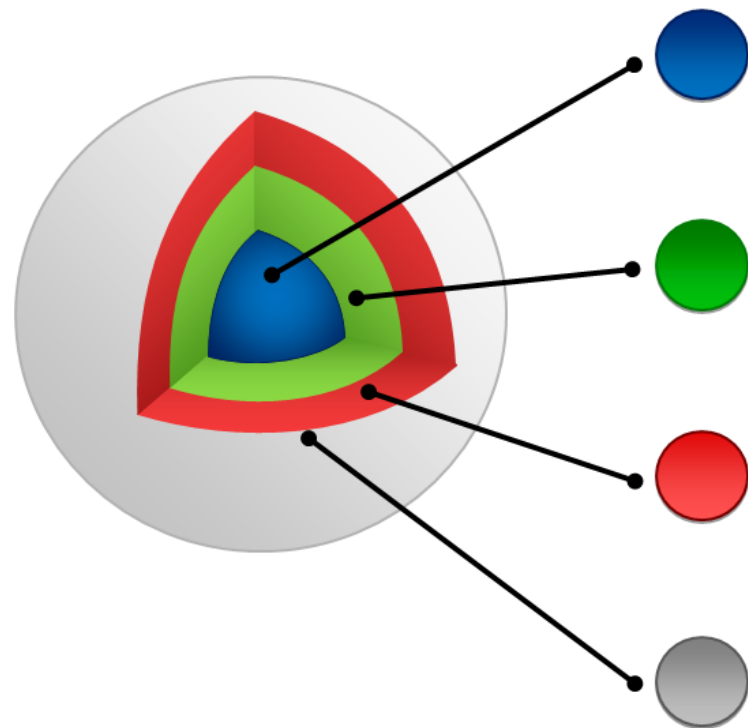
- FAIR Data is a means to reach an end
- The end goal is NOT FAIR Data but better ANALYTICS, more EFFICIENCY and IMPACT on research ROI
- The Internet of FAIR Data and Services is the 'vehicle'
- The tooling supports a data life cycle process
- The FAIRification process requires professional services

Organizations in the process of going FAIR



- Many organizations (20+) have participated in FAIR BYODs and trainings
- Several academic institutions and funders have started or are considering GO FAIR Readiness programs
- Several companies have started or are considering the GO FAIR Readiness program

FAIR Digital Objects and the FAIR Ecosystem



DIGITAL OBJECT

Data, code and other research resources

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique identifiers (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

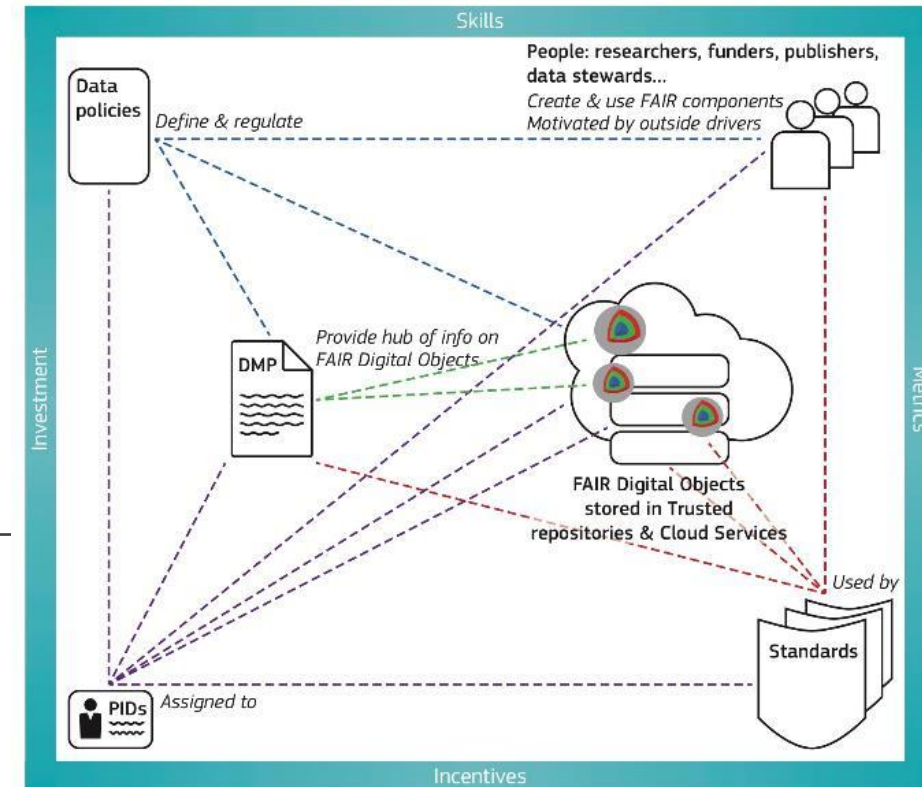
Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

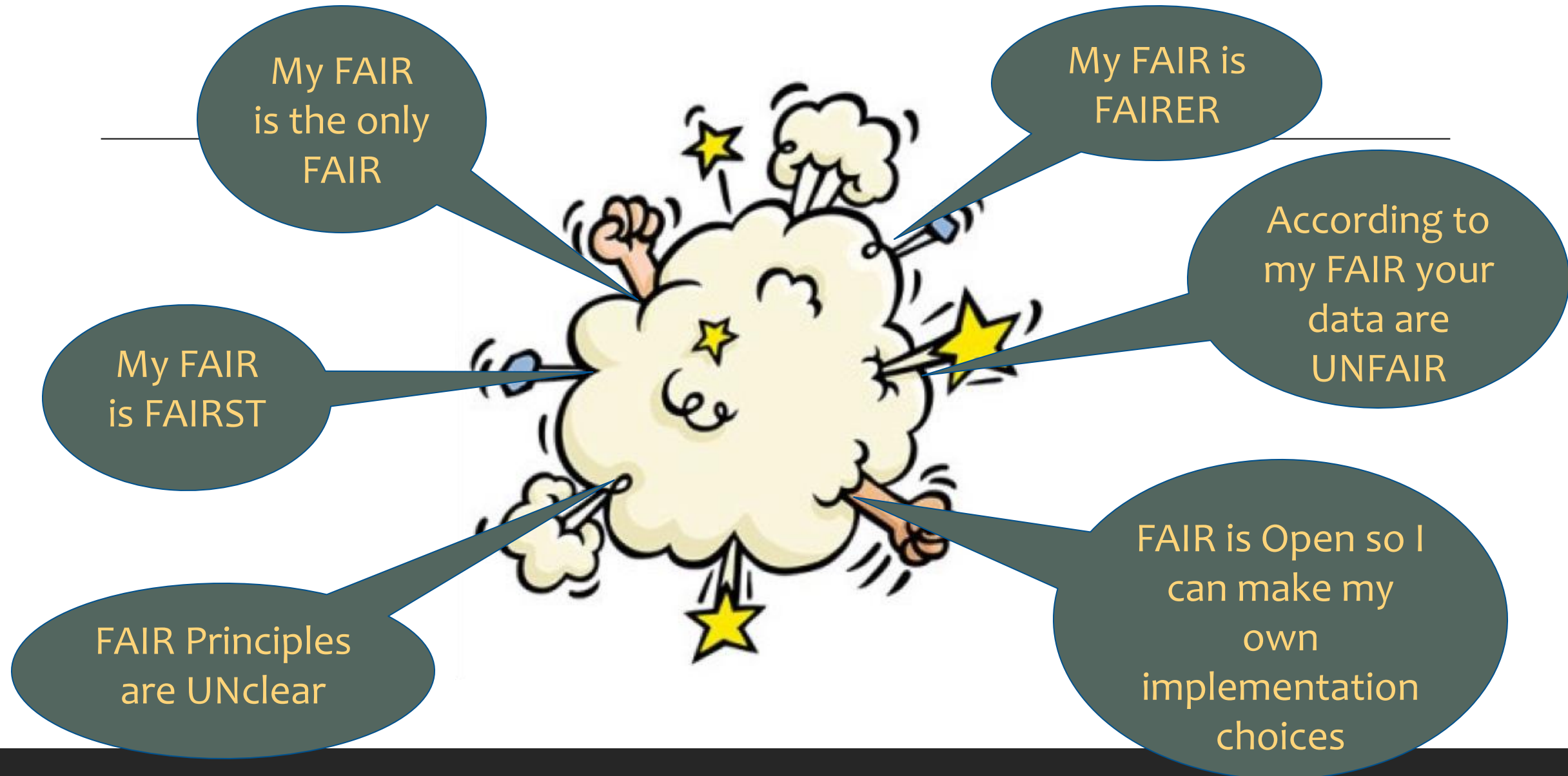
METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.



The FAIR metrics: up for a good fight?



FAIR Tooling Ecosystem

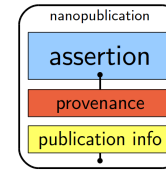
Plan

Create

Publish

Find

Evaluate



Ontology
Modeling

Provenance & Metadata
Management

Metadata
Registry

All tools are currently professorware
Need to be turned into professionalware

FAIR Tooling Ecosystem

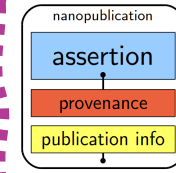
Plan

Create

Publish

Find

Evaluate



Ontology
Modeling

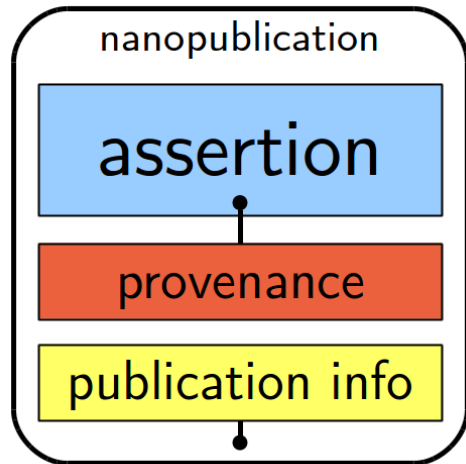
Provenance & Metadata
Management

Metadata
Registry

All tools are currently professorware
Need to be turned into professionalware

FAIR Principles now published as a Trusty Nanopublication

A nanopublication has three basic elements:



1.Assertion: The assertion is the main content of a nanopublication in the form of an small atomic unit of information

2.Provenance: This part describes how the assertion above came to be. This can include the scientific methods that were used to generate the assertion, for example a reference to the kind of study that was performed and its parameters.

3.Publication Info: This part contains metadata about the nanopublication as a whole, such as when and by whom it was created and the license terms for its reuse.

Nanopublications are implemented in the language RDF and come with an evolving ecosystem of tools and systems.

```

@prefix this: <http://purl.org/np/RA6nM9ZNUoc3HU8ODzokamR-LPySCqDR1zJfutGhcwrqg> .
@prefix sub: <http://purl.org/np/RA6nM9ZNUoc3HU8ODzokamR-LPySCqDR1zJfutGhcwrqg#> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix hycl: <http://purl.org/petapico/o/hycl#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix npx: <http://purl.org/nanopub/x/> .

```

```

sub:Head {
  this: np:hasAssertion sub:assertion ;
  np:hasProvenance sub:provenance ;
  np:hasPublicationInfo sub:pubInfo ;
  a np:Nanopublication .
}

```

```

sub:assertion {
  sub:mystatement a hycl:Statement ;
  rdfs:label "Fundamentos de Data Science" .
}

```

```

sub:provenance {
  sub:assertion prov:generatedAtTime "2021-04-07T17:02:22.048026"^^xsd:dateTime ;
  prov:wasAttributedTo <https://orcid.org/0000-0002-0792-8157> .
  <https://orcid.org/0000-0002-0792-8157> hycl:claims sub:mystatement .
}

```

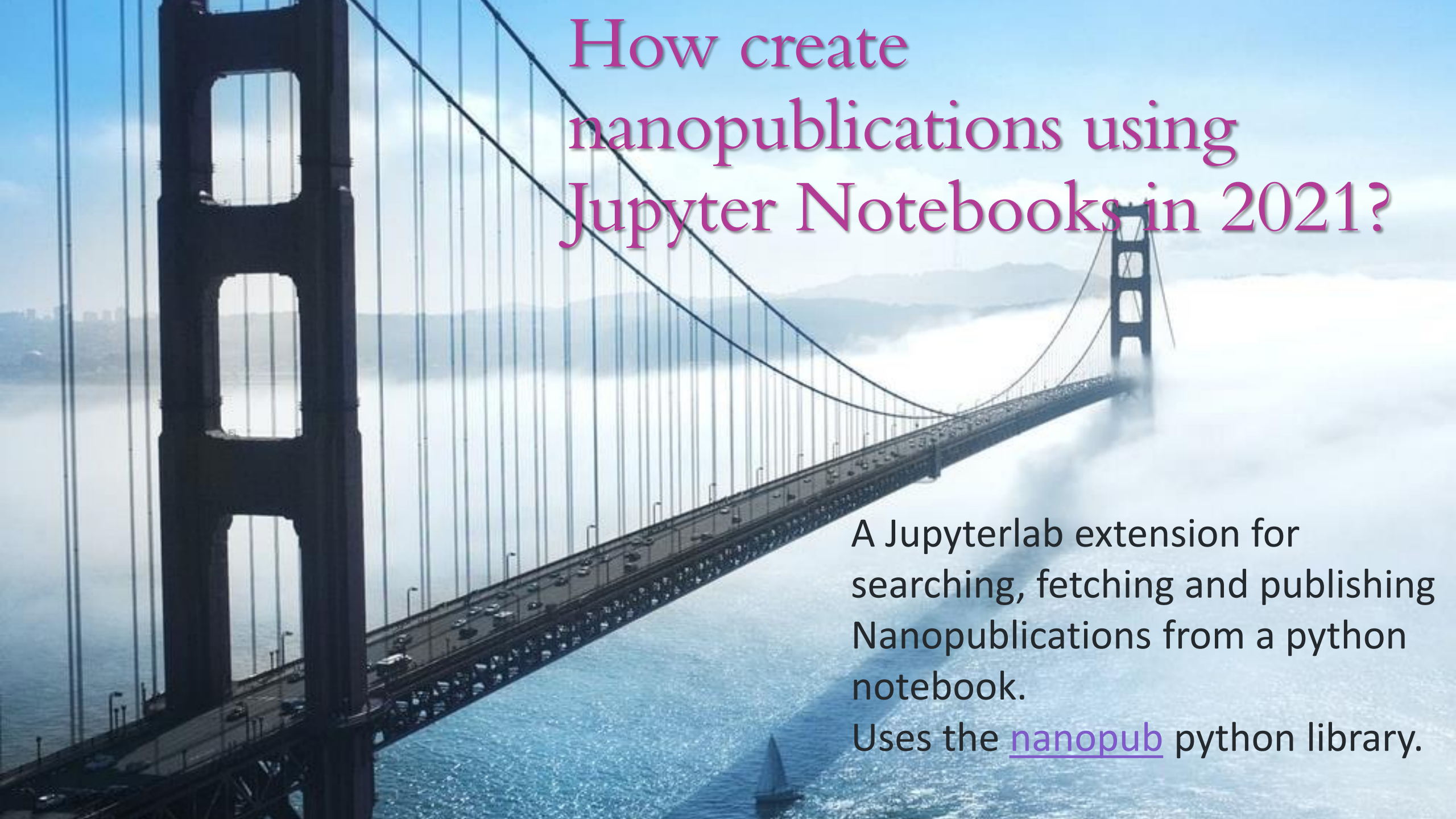
```

sub:pubInfo {
  sub:sig npx:hasAlgorithm "RSA" ;
  npx:hasPublicKey
"MIIGfMA0GCSqGSIb3DQEBAQUAA4GNADCBiQKBgQCqDYQKXdt0wLyWC9nsBpURQCYDd7uMZqIHewWZzcXHy7qT3mqKgWuW2dmEQV+rSr+OjNnzKDKEnR/h25m0IQcgJDEy7aa0gwokN+9Gwf0pWm1HyneYb28Qa0UonyXpR1y5ilQ3kblmH2PMGJEdifCp0uk70NeRls05IGIB1ZkNmWIDAQAB"
;
  npx:hasSignature "E2LSOVouQVbdo/gbi7aCewK18uJJLmf3A86mhGGPVx+2lCng2kebiSGwdDgq/HoR4z/whZWSNBA+F7sK7xdgeXZBxnDVVzPElsogm4J/CQJGKBORvDT01HRb/p8PKCc6gj9PxrrfbnShp9D2sBiXI1hkrCcv3QnBLL4dtlG+h+0=" ;
  npx:hasSignatureTarget this: .
  this: prov:generatedAtTime "2021-04-07T17:02:22.048026"^^xsd:dateTime ;
  prov:wasAttributedTo <https://orcid.org/0000-0002-0792-8157> .
}

```



Nanopublications: A simple way to communicate complex science!!!

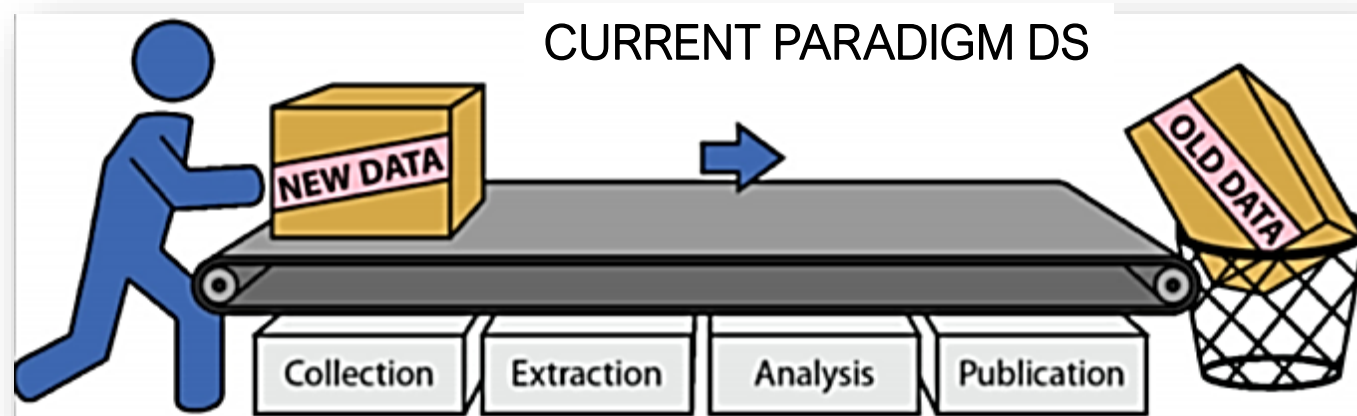
A photograph of the Golden Gate Bridge in San Francisco, viewed from a low angle looking down the length of the bridge towards the other side. The bridge's iconic orange-red towers and suspension cables are prominent against a blue sky with light clouds. The water below is a deep blue, and a small sailboat is visible in the distance. The overall scene is bright and clear.

How create nanopublications using Jupyter Notebooks in 2021?

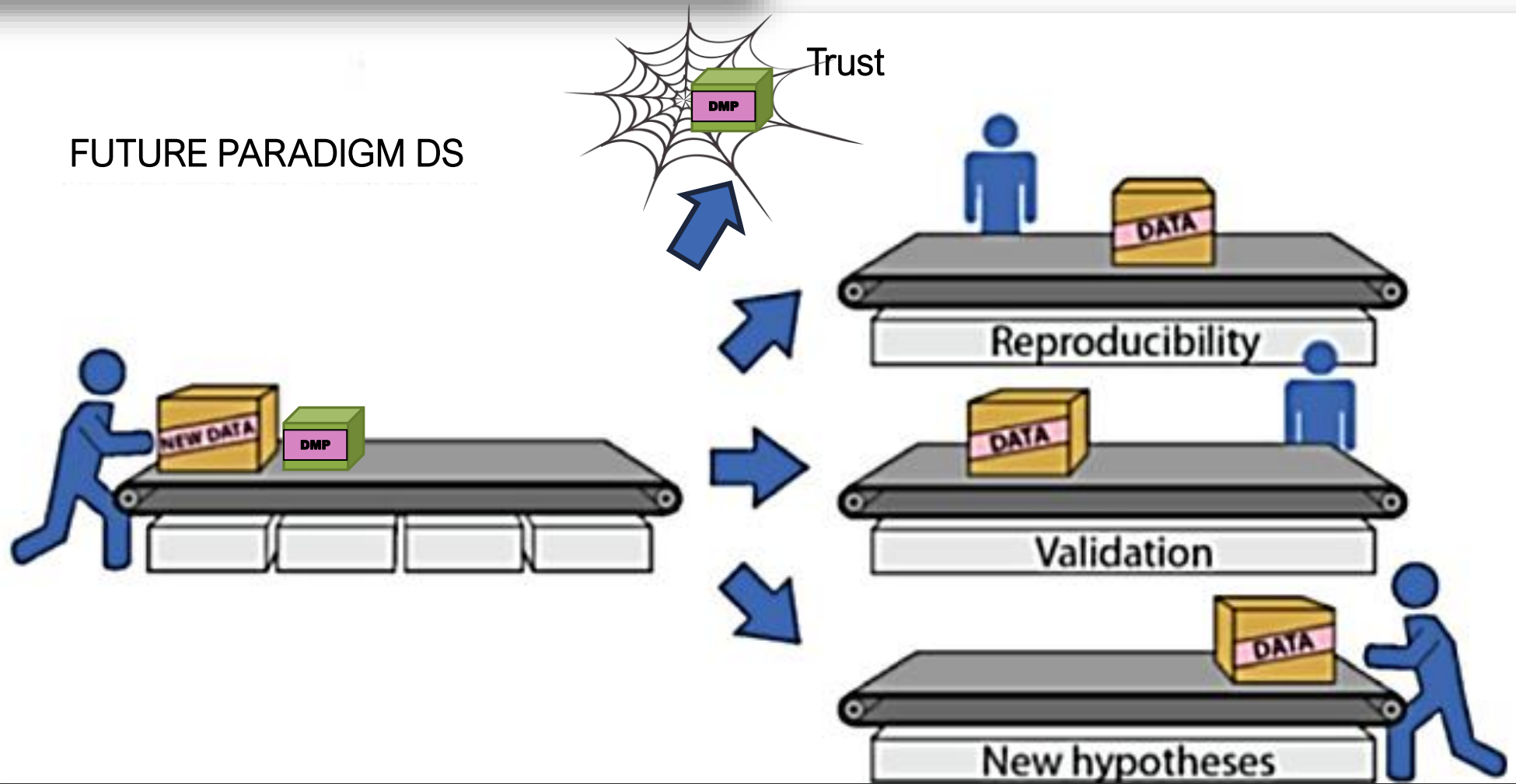
A Jupyterlab extension for
searching, fetching and publishing
Nanopublications from a python
notebook.

Uses the [nanopub](#) python library.

CURRENT PARADIGM DS



FUTURE PARADIGM DS



What is a Data Management Plan?

A brief plan (~2 pages) written at the start of your project to define:

1. how your data will be created?
2. how it will be documented?
3. who will access it?
4. where it will be stored?
5. who will back it up?
6. whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever you're creating NEW data.



A document that describes what you will do with your data **during** and **after** you complete your research



Why develop a DMP?

To help you manage your datasets → Ensure that you and others will be able to understand and use data in future → **Reproducible Data Management**

To make informed decisions so you don't have to figure out things as you go

To anticipate and avoid problems, like:

- data loss,
- lack of trust,
- transparency,
- authorship,
- patenting wars, etc..

To save time, satisfy funding agencies and make your life easier! (e.g. H2020, NSF, FAPESP, Tecnopolo project)

Common themes in DMPs

1. Description of what data to be collected / created / produced
(*i.e.* content, type, file formats, provenance, volume, who create..)
2. Standards / methodologies for data collection & management
(*i.e.* Metadata and documentation and standards used, how organize..)
3. Ethics and Intellectual Property implications
(highlight any restrictions on data sharing *e.g.* ownership, embargoes, confidentiality)
4. Plans/policies for data sharing, reuse and access (or licensing)
(*i.e.* how, when, to whom)
5. Strategy for long-term preservation
(which data need to keep, who will keep, where to keep, (FAIR) repositories...)

How develop a DMP?

Use a DMPs framework!

They ensure you address all areas of data management


DMPs do not check or validate your answers!





Top 7 DMP Tools


1. [Adobe DMP](#)
 2. [Neustar IDMP](#)
 3. [Oracle DMP](#)
 4. [Salesforce DMP](#)
 5. [Lotame](#)
 6. [Mapp Acquire](#)
 7. [Nielsen DMP](#)
-


<https://ds-wizard.org/>


 DS Wizard


 Knowledge Model Editor


 Knowledge Models

 Projects

 Templates

 Storage Costs Evaluator

 Help >

 Sergio Serra >

<< Collapse sidebar

This is only **demonstration** instance intended for testing DSW features. For serious work, please use **Researchers** or your institutional instance. **All data are periodically deleted** (excluding user accounts) from this instance (every 1st of month at 2AM UTC).

Welcome to the DS Wizard!



Choose a Knowledge Model
suitable for your Project.



Fill in the Questionnaire,
chapter by chapter.



Get your Data Management Plan
ready to be submitted.

Start planning

Researchers from these institutions are using DSW



<https://ds-wizard.org/get-started.html>



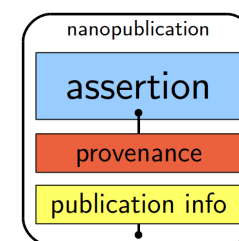


<https://ds-wizard.org/get-started.html>

Hands on...



NOTEBOOK:
NANOPUBLICATIONS



Ten Simple Rules for Creating a Good Data Management Plan

William K. Michener Published: October 22, 2015 • <https://doi.org/10.1371/journal.pcbi.1004525>

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve , Anton Nekrutenko, James Taylor, Eivind HovigPublished: October 24, 2013 • <https://doi.org/10.1371/journal.pcbi.1003285>

Good enough practices in scientific computing

Greg Wilson , Jennifer Bryan , Karen Cranston , Justin Kitzes , Lex Nederbragt , Tracy K. Teal Published: June 22, 2017 • <https://doi.org/10.1371/journal.pcbi.1005510>

References

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525>

<https://elifesciences.org/labs/d42fe2b9/integrating-binder-and-stencila-the-building-blocks-to-increased-open-communication-and-transparency>

<https://ds-wizard.org>

<https://mybinder.org/>

<https://stenci.la/>

<https://www.mysciencework.com/>

<https://fapesp.br/gestaodedados>

arXiv.org > cs > arXiv:1810.08055

Search...

Help | Advanced

Computer Science > Other Computer Science

[Submitted on 13 Oct 2018]

Ten Simple Rules for Reproducible Research in Jupyter Notebooks

Adam Rule, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, Mai H. Nguyen, Sara Brin Rosenthal, Fernando Pérez, Peter W. Rose

OPINION ARTICLE

Related to other papers in this special issue	25 (q246); 26 (q257); 27 (q264)
Addressing FAIR principles	F, A, I, R

GO FAIR Brazil: A Challenge for Brazilian Data Science

Luana Sales¹, Patrícia Henning^{2*}, Viviane Veiga³, Maira Murieta Costa⁴, Luís Fernando Sayão⁵, Luiz Olavo Bonino da Silva Santos⁶ & Luís Ferreira Pinheiro⁷¹Instituto Brasileiro em Ciência e Tecnologia, Rio de Janeiro - RJ, 22290-140, Brazil²Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro - RJ, 22290-240, Brazil³Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasília, DF, CEP 70002-900, Brazil⁴Ministério da Ciência, Tecnologia, Inovação e Comunicação, Esplanada dos Ministérios, Bloco F - Brasília, DF, CEP 70002-900, Brazil⁵Comissão Nacional de Energia Nuclear, Rua Gal. Senechal, s/nº, Barro Preto, Belo Horizonte, Minas Gerais, CEP 31270-900 - Belo Horizonte, Brazil⁶GO FAIR International Support & Coordination Office (GIFCO), London E13 4AA, The Netherlands⁷University of Twente, Enschede 7522 NB, The Netherlands**Keywords:** FAIR principles; GO FAIR; GO FAIR Brazil; Open Science; Research dataCitation: Sales L, Henning P, Veiga V, Murieta Costa M, Sayão LF, Bonino da Silva Santos LO, et al. (2018) GO FAIR Brazil: A challenge for Brazilian data science. *Data Intelligence* 2(2020): 238-245. doi: 10.1162/dint_a_00046

ABSTRACT

The FAIR principles, an acronym for Findable, Accessible, Interoperable and Reusable, are recognised worldwide as key elements for good practice in all data management processes. To understand how the Brazilian scientific community is adhering to these principles, this article reports Brazilian adherence to the GO FAIR initiative through the creation of the GO FAIR Brazil Office and the manner in which they create their implementation networks. To contextualise this understanding, we provide a brief presentation of open data policies in Brazilian research and government, and finally, we describe a model that has been adopted

* Corresponding author: Patrícia Henning (E-mail: henningpatricia@gmail.com; ORCID: 0000-0001-0739-6442).