- Science Ouverte et reproductible : éléments de contexte et pistes de mise en oeuvre

Qui suis-je ?

Qui suis-je ?

- PhD Plant Physio & Biochemistry
- Post-Doc in Genetics
- Researcher
- Montpellier

Qui suis-je ?

- PhD Plant Physio & Biochemistry
- Post-Doc in Genetics
- Researcher
- Montpellier

- Biochemistry
- Biophysics
- 3D Modeling
- Bioinformatics
- Bioinformatics platform

Mes activités

National Research Institute for Agriculture, Food and Environment
***Data Steward* : Plant Biology and Breeding Division**
1200 people & 22 Research Units and 19 Plant Facilities

https://orcid.org/0000-0003-4234-1172

Mes activités

National Research Institute for Agriculture, Food and Environment
***Data Steward* : Plant Biology and Breeding Division**
1200 people & 22 Research Units and 19 Plant Facilities

https://orcid.org/0000-0003-4234-1172

Mes activités

National Research Institute for Agriculture, Food and Environment
***Data Steward* : Plant Biology and Breeding Division**
1200 people & 22 Research Units and 19 Plant Facilities

IFB is the National Bioinformatics Infrastructure that provides support, services, training and developments for the life sciences communities.
Chief Data Officer

https://orcid.org/0000-0003-4234-1172

Mes activités

National Research Institute for Agriculture, Food and Environment
***Data Steward* : Plant Biology and Breeding Division**
1200 people & 22 Research Units and 19 Plant Facilities

IFB is the National Bioinformatics Infrastructure that provides support, services, training and developments for the life sciences communities.
Chief Data Officer

https://orcid.org/0000-0003-4234-1172

Mes activités

National Research Institute for Agriculture, Food and Environment
***Data Steward* : Plant Biology and Breeding Division**
1200 people & 22 Research Units and 19 Plant Facilities

IFB is the National Bioinformatics Infrastructure that provides support, services, training and developments for the life sciences communities.
Chief Data Officer

ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research.

https://orcid.org/0000-0003-4234-1172

Un peu de lecture

Un peu de lecture

Hamilton *et al. BMC Medicine* (2022) 20:438
https://doi.org/10.1186/s12916-022-02644-2

BMC Medicine

RESEARCH ARTICLE Open Access

How often do cancer researchers make their data and code available and what factors are associated with sharing?

Daniel G. Hamilton[1,2*], Matthew J. Page[3], Sue Finch[4], Sarah Everitt[5] and Fiona Fidler[1,6]

Un peu de lecture

Hamilton *et al. BMC Medicine* (2022) 20:438
https://doi.org/10.1186/s12916-022-02644-2

BMC Medicine

RESEARCH ARTICLE Open Access

How often do cancer researchers make their data and code available and what factors are associated with sharing?

Daniel G. Hamilton[1,2*], Matthew J. Page[3], Sue Finch[4], Sarah Everitt[5] and Fiona Fidler[1,6]

Un peu de lecture

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data

European Commission
Directorate-General for Research and Innovation
Directorate A — Policy Development and Coordination
Unit A.2 — Open Data Policy and Science Cloud
Contact Athanasios Karalopoulos
E-mail Athanasios.Karalopoulos@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Hamilton *et al. BMC Medicine* (2022) 20:438
https://doi.org/10.1186/s12916-022-02644-2

BMC Medicine

RESEARCH ARTICLE Open Access

How often do cancer researchers make their data and code available and what factors are associated with sharing?

Daniel G. Hamilton[1,2*], Matthew J. Page[3], Sue Finch[4], Sarah Everitt[5] and Fiona Fidler[1,6]

Un peu de lecture

Hamilton *et al. BMC Medicine* (2022) 20:438
https://doi.org/10.1186/s12916-022-02644-2

BMC Medicine

RESEARCH ARTICLE Open Access

How often do cancer researchers make their data and code available and what factors are associated with sharing?

Daniel G. Hamilton[1,2*], Matthew J. Page[3], Sue Finch[4], Sarah Everitt[5] and Fiona Fidler[1,6]

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data
European Commission
Directorate-General for Research and Innovation
Directorate A — Policy Development and Coordination
Unit A.2 — Open Data Policy and Science Cloud
Contact Athanasios Karalopoulos
E-mail Athanasios.Karalopoulos@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Un peu de lecture

Besançon *et al. BMC Medical Research Methodology* (2021) 21:117
https://doi.org/10.1186/s12874-021-01304-y

BMC Medical Research Methodology

COMMENTARY Open Access

Open science saves lives: lessons from the COVID-19 pandemic

Lonni Besançon[1,2*], Nathan Peiffer-Smadja[3,4], Corentin Segalas[5], Haiting Jiang[6], Paola Masuzzo[7], Cooper Smout[7], Eric Billy[8], Maxime Deforet[9] and Clémence Leyrat[5,10]

Hamilton *et al. BMC M*
https://doi.org/10.118

RESEARCH ARTICLE Open Access

How often do cancer researchers make their data and code available and what factors are associated with sharing?

Daniel G. Hamilton[1,2*], Matthew J. Page[3], Sue Finch[4], Sarah Everitt[5] and Fiona Fidler[1,6]

European Commission

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data
European Commission
Directorate-General for Research and Innovation
Directorate A — Policy Development and Coordination
Unit A.2 — Open Data Policy and Science Cloud
Contact Athanasios Karalopoulos
E-mail Athanasios.Karalopoulos@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Plan

- Context (what happened ?)
- Research Data (what do I have to do ?)
  - Findable
  - Accessible
  - Interoperable
  - Reusable
- Data Quality
- Data Analysis
- What's next ?
- A Success Story
- A take away list

Pourquoi parler de Science Ouverte maintenant ?

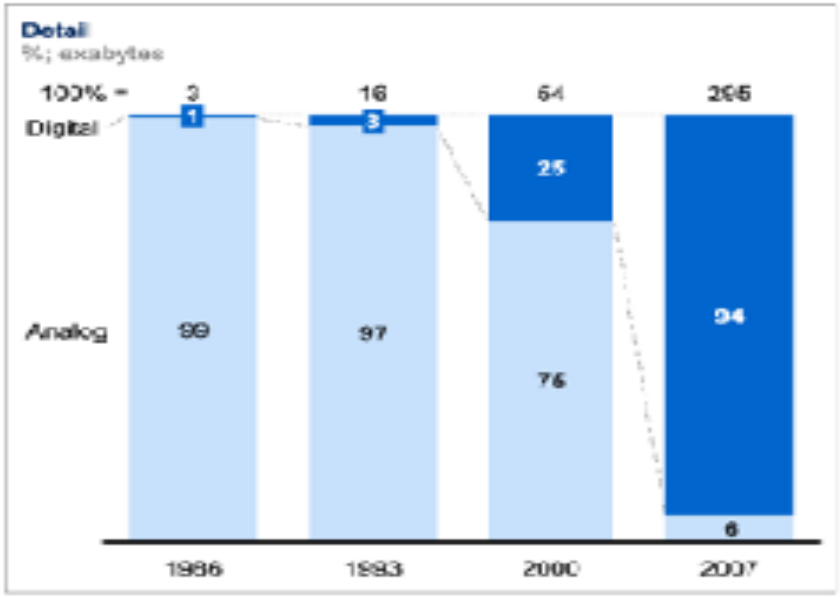
On fait de la Science depuis des décennies 🤷 !

Que c’est il passé ?

Que c’est il passé ?

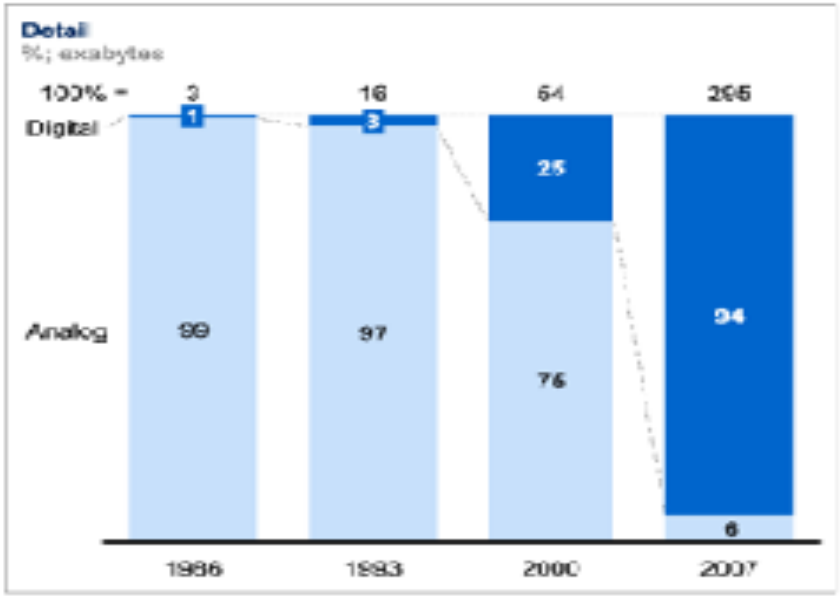
Digital disruption, an abrupt shift

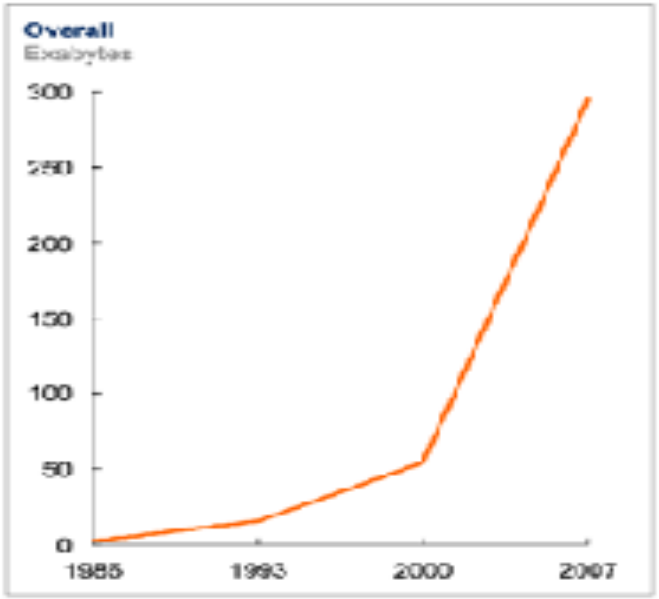
Digital disruption, an abrupt shift



From : The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 2011;332.

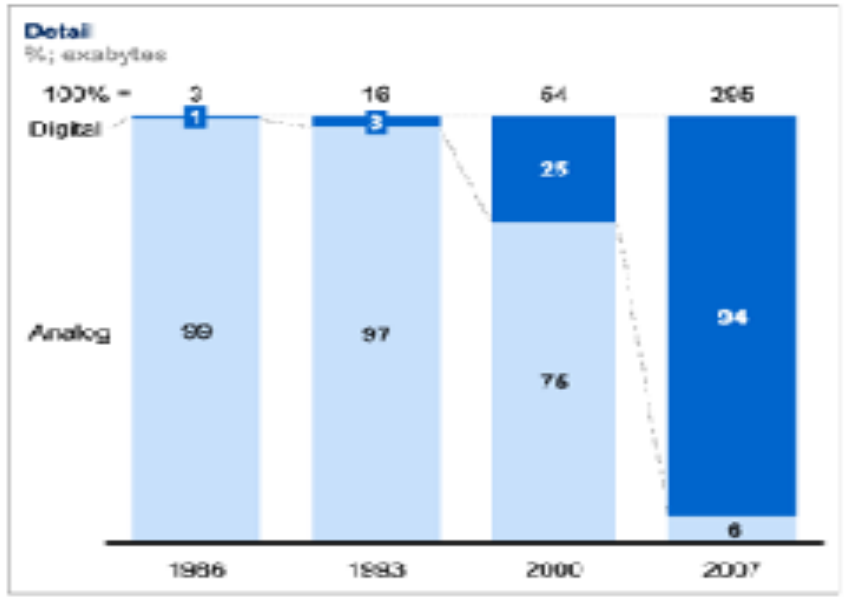
Digital disruption, an abrupt shift

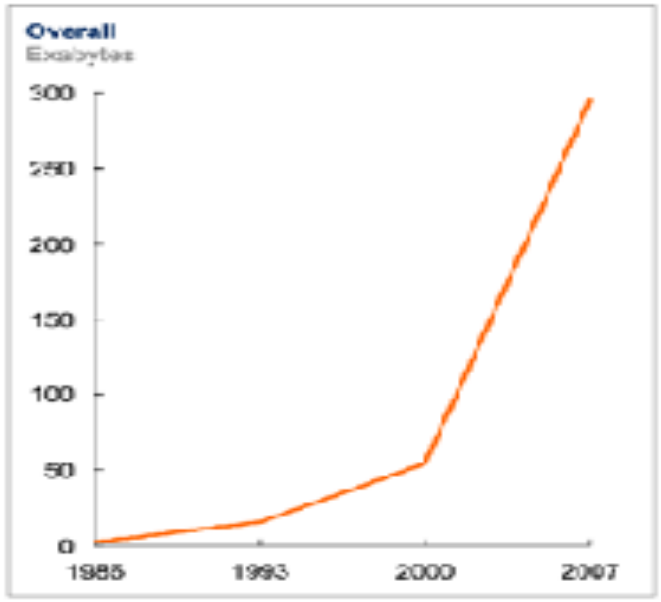




From : The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 2011;332.

Digital disruption, an abrupt shift





From : The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 2011;332.

… Shuffling the deck

… Shuffling the deck

- First cab company has no cab (Uber)

… Shuffling the deck

- First cab company has no cab (Uber)
- First accommodation company owns no room (AirBnB)

… Shuffling the deck

- First cab company has no cab (Uber)
- First accommodation company owns no room (AirBnB)
- First voice conversation provider owns no switchboard (Skype)

… Shuffling the deck

- First cab company has no cab (Uber)
- First accommodation company owns no room (AirBnB)
- First voice conversation provider owns no switchboard (Skype)
- First content provider does not produce content (Facebook)

… Shuffling the deck

- First cab company has no cab (Uber)
- First accommodation company owns no room (AirBnB)
- First voice conversation provider owns no switchboard (Skype)
- First content provider does not produce content (Facebook)
- First movies (and series) provider owns no theater (Netflix)

… Shuffling the deck

- First cab company has no cab (Uber)
- First accommodation company owns no room (AirBnB)
- First voice conversation provider owns no switchboard (Skype)
- First content provider does not produce content (Facebook)
- First movies (and series) provider owns no theater (Netflix)

The data deluge in Science

High-throughput techniques drive down costs and boost data production : The Human Genome

1990 = **13 years** & **3 Billions** $

2015 = **Hours** & **1000** $

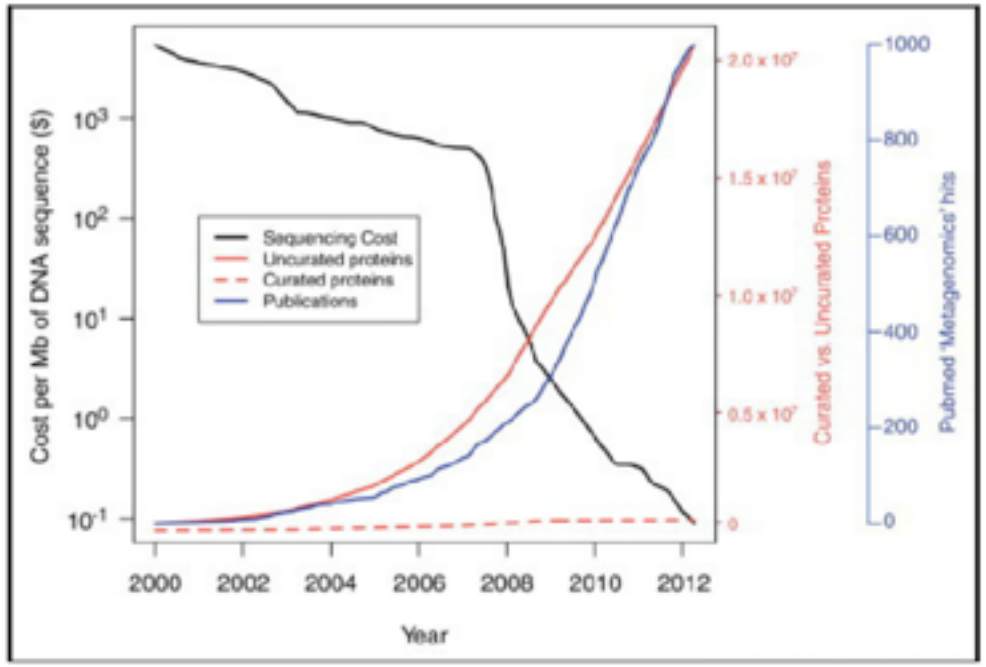
From : Using modern plant breeding to improve the nutritional and technological qualities of oil crops DOI: 10.1051/ocl/2014038

The data deluge in Science

High-throughput techniques drive down costs and boost data production : The Human Genome

1990 = **13 years** & **3 Billions** $

2015 = **Hours** & **1000** $



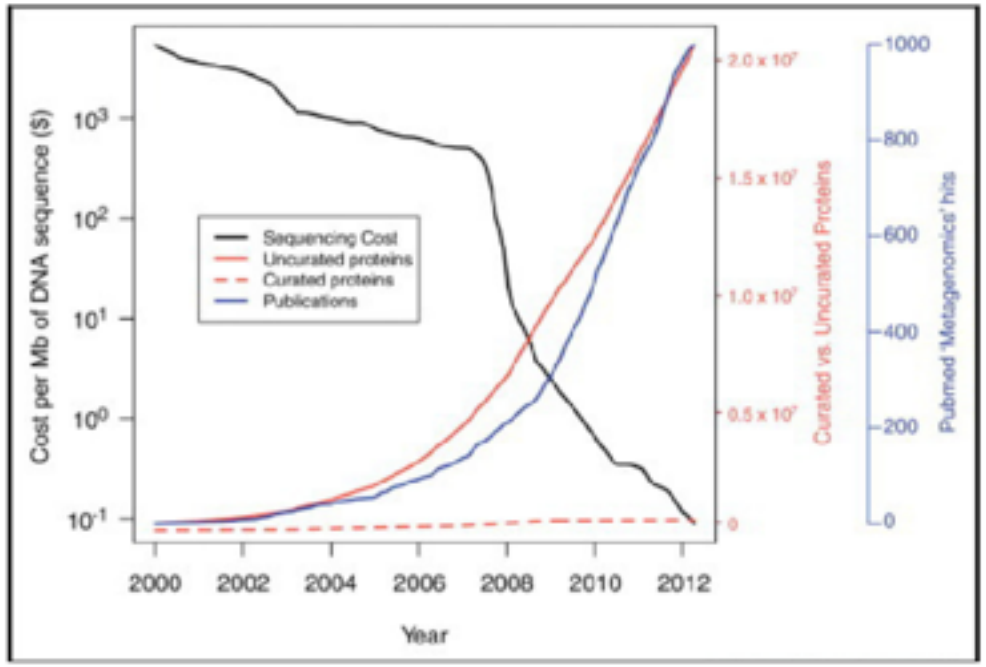
From : Using modern plant breeding to improve the nutritional and technological qualities of oil crops DOI: 10.1051/ocl/2014038

The data deluge in Science

High-throughput techniques drive down costs and boost data production : The Human Genome

1990 = **13 years** & **3 Billions** $

2015 = **Hours** & **1000** $



- The amount of data to be stored and analyzed is rocketing
- Analysis performance declines

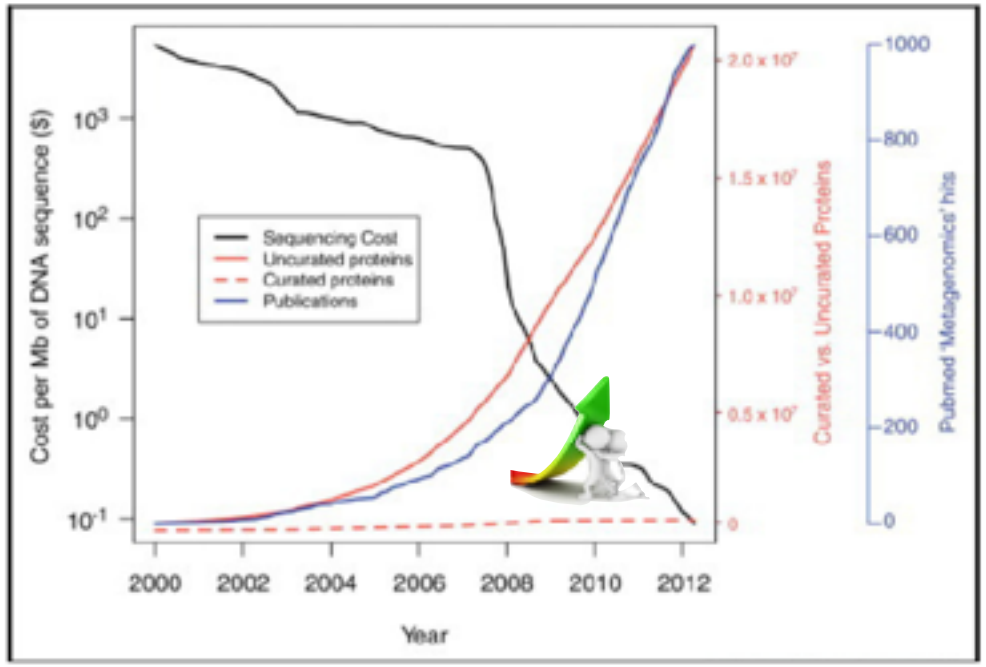
From : Using modern plant breeding to improve the nutritional and technological qualities of oil crops DOI: 10.1051/ocl/2014038

The data deluge in Science

High-throughput techniques drive down costs and boost data production : The Human Genome

1990 = **13 years** & **3 Billions** $

2015 = **Hours** & **1000** $



- The amount of data to be stored and analyzed is rocketing
- Analysis performance declines

From : Using modern plant breeding to improve the nutritional and technological qualities of oil crops DOI: 10.1051/ocl/2014038

A new way of « doing science »

Before

1 Designing the experiment
2 Collecting results
3 Analyzing the results

A new way of « doing science »

Before

1 Designing the experiment
2 Collecting results
3 Analyzing the results

Now

1 Massive data generation
2 Organise (store, document, annotate)
3 Analyse (extract information)
4 Disseminate information

A change of scale that modifies the how knowledge is produced

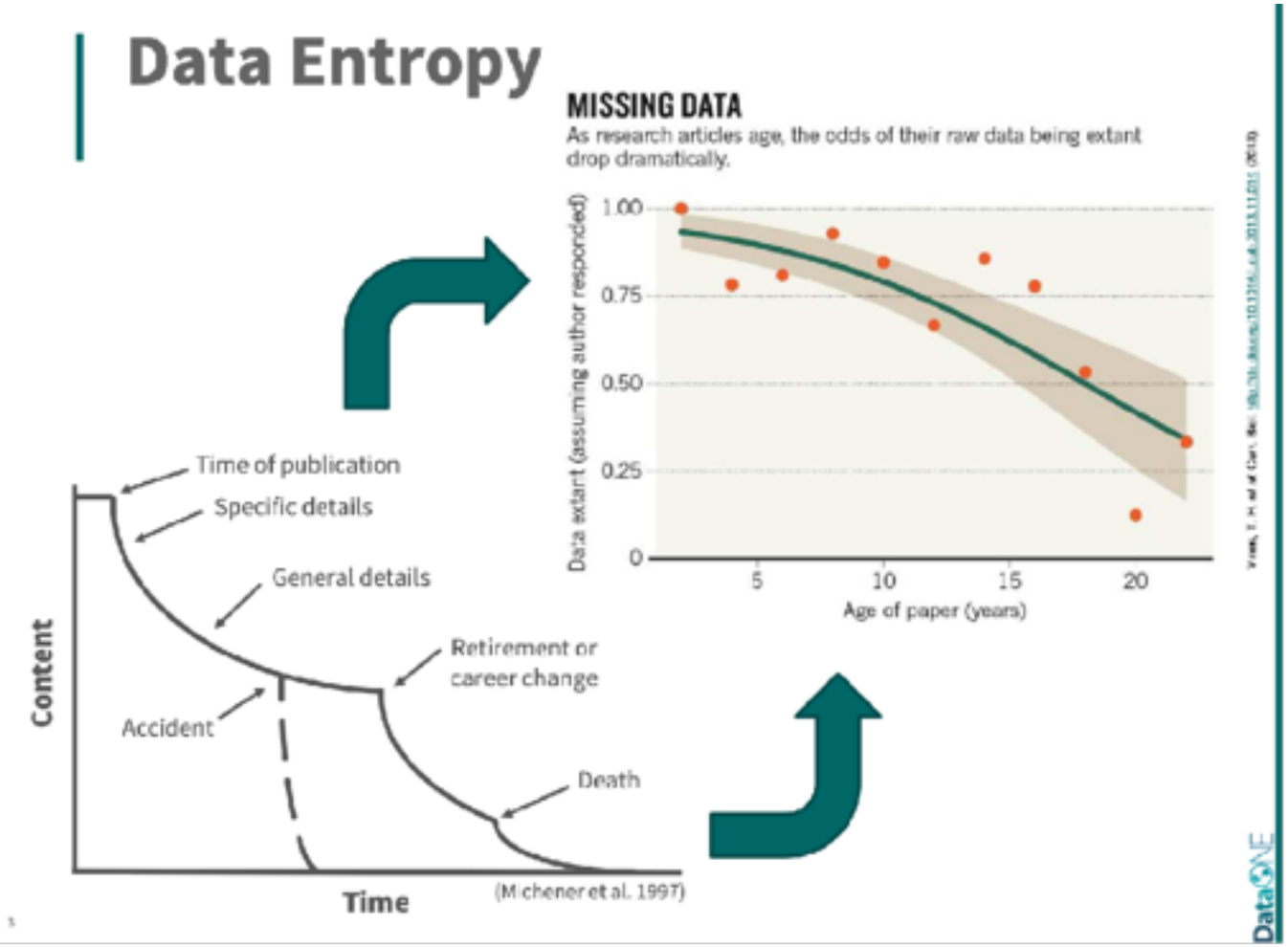
A change of scale that modifies the how knowledge is produced

THE HUMAN GENOME

The Sequence of the Human Genome

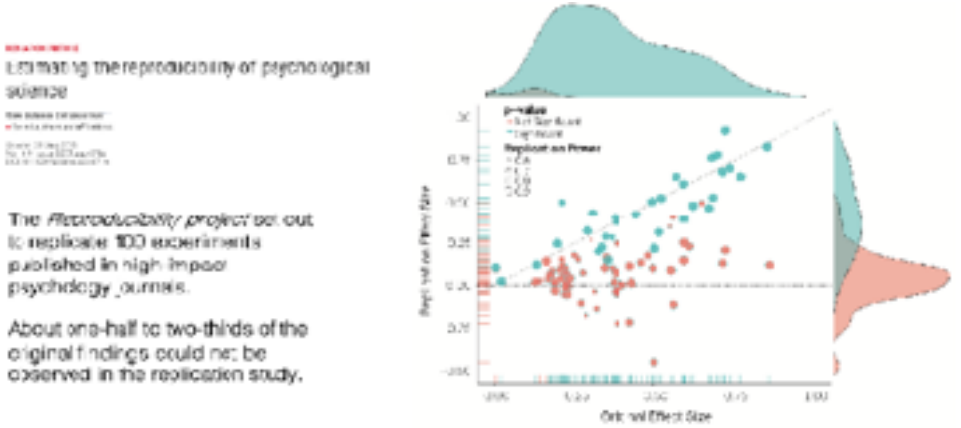
J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Ganglaharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[2] Samuel Broder,[3] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[1] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[1,2] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[13] Suzanna Baumhueter,[13] Gene Spier,[13] Christine Carter,[14] Anibal Cravchik,[14] Trevor Woodage,[14] Feroze Ali,[14] Huijin An,[14] Aderonke Awe,[14] Danita Baldwin,[14] Holly Baden,[14] Mary Bainbridge,[14] Ian Barrow,[14] Karen Beeson,[14] Dana Busam,[14] Amy Carver,[14] Angela Center,[14] Ming Lai Cheng,[14] Liz Curry,[14] Steve Dandnher,[14] Lionel Davenport,[14] Raymond Desilets,[14] Susanne Dietz,[14] Kristina Dodson,[14] Lisa Doup,[14] Steven Ferriera,[14] Neha Garg,[14] Andres Gluecksmann,[14] Brit Hart,[14] Jason Haynes,[14] Charles Haynes,[14] Cheryl Heiner,[14] Suzanne Hobbs,[14] Damon Hostin,[14] Jarrett Houck,[14] Timothy Howland,[14] Chinyere Ibegwam,[14] Jeffery Johnson,[14] Francis Kalush,[14] Lesley Kline,[14] Shashi Koduru,[14] Amy Love,[14] Felecia Mann,[14] David May,[14] Steven McCawley,[14] Tina McIntosh,[14] Ivy McMullen,[14] Mee Moy,[14] Linda Moy,[14] Brian Murphy,[14] Keith Nelson,[14] Cynthia Pfannkoch,[14] Eric Pratts,[14] Vinita Puri,[14] Hina Qureshi,[14] Matthew Reardon,[14] Robert Rodriguez,[14] Yu-Hui Rogers,[14] Deanna Romblad,[14] Bob Ruhfel,[14] Richard Scott,[14] Cynthia Sitter,[14] Michelle Smallwood,[14] Erin Stewart,[14] Renee Strong,[14] Ellen Suh,[14] Reginald Thomas,[14] Ni Ni Tint,[14] Sukyan Tse,[14] Claire Vech,[14] Gary Wang,[14] Jeremy Wetter,[14] Sherita Williams,[14] Monica Williams,[14] Sandra Windsor,[14] Emily Winn-Deen,[14] Kerfalla Wolfe,[14] Jayshree Zaveri,[14] Karena Zaveri,[14] Josep F. Abril,[15,16] Roderic Guigó,[15,16] Michael J. Campbell,[17] Kimmen V. Sjolander,[17] Brian Karlak,[17] Anish Kejariwal,[17] Huaiyu Mi,[17] Betty Lazareva,[17] Thomas Hatton,[17] Apurva Narechania,[17] Karen Diemer,[17] Anuradha Muruganujan,[17] Nan Guo,[17] Shinji Sato,[17] Vineet Bafna,[18] Sorin Istrail,[18] Russ Lippert,[18] Russell Schwartz,[18] Brian Walenz,[18] Shibu Yooseph,[18] David Allen,[18] Anand Basu,[18] James Baxendale,[18] Louis Blick,[18] Marcelo Caminha,[18] John Carnes-Stine,[18] Parris Caulk,[18] Yen-Hui Chiang,[18] My Coyne,[18] Carl Dahlke,[18] Anne Deslattes Mays,[18] Maria Dombroski,[18] Michael Donnelly,[18] Dale Ely,[18] Shiva Esparham,[18] Carl Fosler,[18] Harold Gire,[18] Stephen Glanowski,[18] Kenneth Glasser,[18] Anna Glodek,[18] Mark Gorokhov,[18] Ken Graham,[18] Barry Gropman,[18] Michael Harris,[18] Jeremy Heil,[18] Scott Henderson,[18] Jeffrey Hoover,[18] Donald Jennings,[18] Catherine Jordan,[18] James Jordan,[18] John Kasha,[18] Leonid Kagan,[18] Cheryl Kraft,[18] Alexander Levitsky,[18] Mark Lewis,[18] Xiangjun Liu,[18] John Lopez,[18] Daniel Ma,[18] William Majoros,[18] Joe McDaniel,[18] Sean Murphy,[18] Matthew Newman,[18] Trung Nguyen,[18] Ngoc Nguyen,[18] Marc Nodell,[18] Sue Pan,[18] Jim Peck,[18] Marshall Peterson,[18] William Rowe,[18] Robert Sanders,[18] John Scott,[18] Michael Simpson,[18] Thomas Smith,[18] Arlan Sprague,[18] Timothy Stockwell,[18] Russell Turner,[18] Eli Vetter,[18] Hai Wang,[18] Mehyar Wen,[18] David Wu,[18] Mitchell Wu,[18] Ashley Xia,[18] Ali Zandieh,[18] Xiaohong Zhu[18]

Damage of time



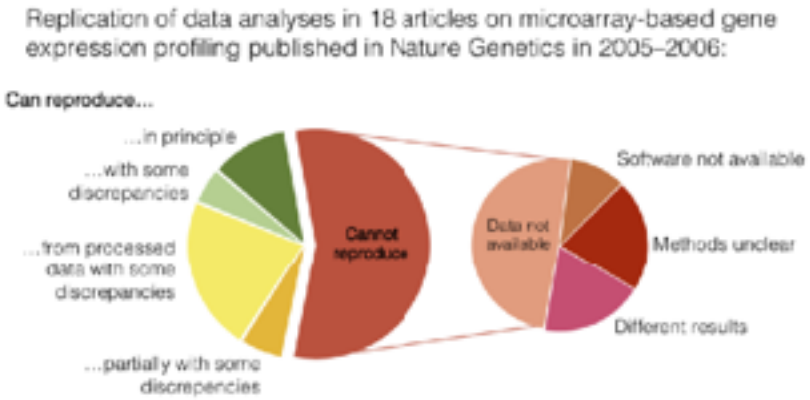
The reproducibility crisis

The reproducibility crisis

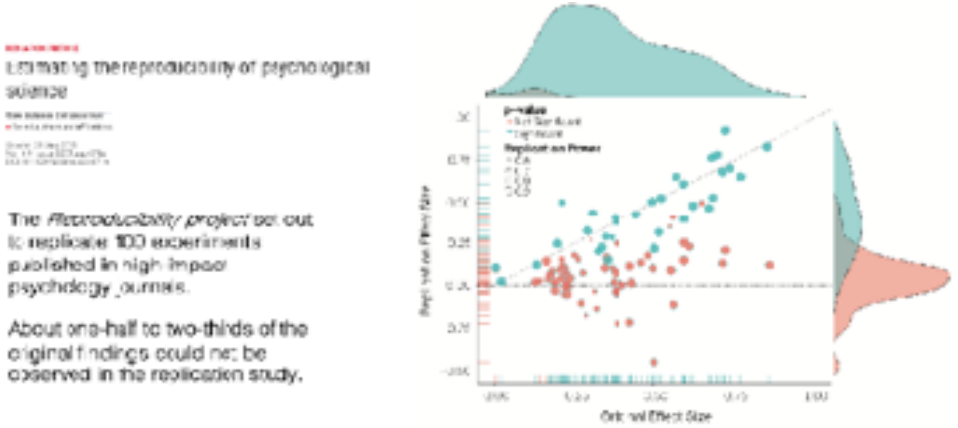


Estimating the reproducibility of psychological science. Science. **2015**;349: aac4716. doi:10.1126/science.aac4716

The reproducibility crisis

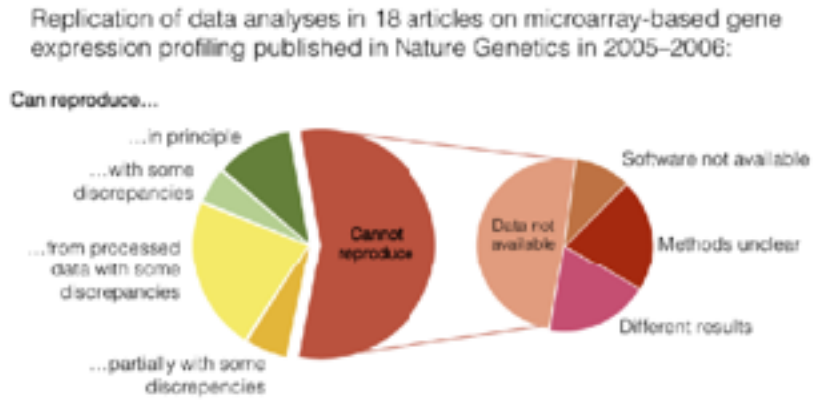


Repeatability of published microarray gene expression analyses.
Nat Genet. **2009**;41: 149–155. doi:10.1038/ng.295

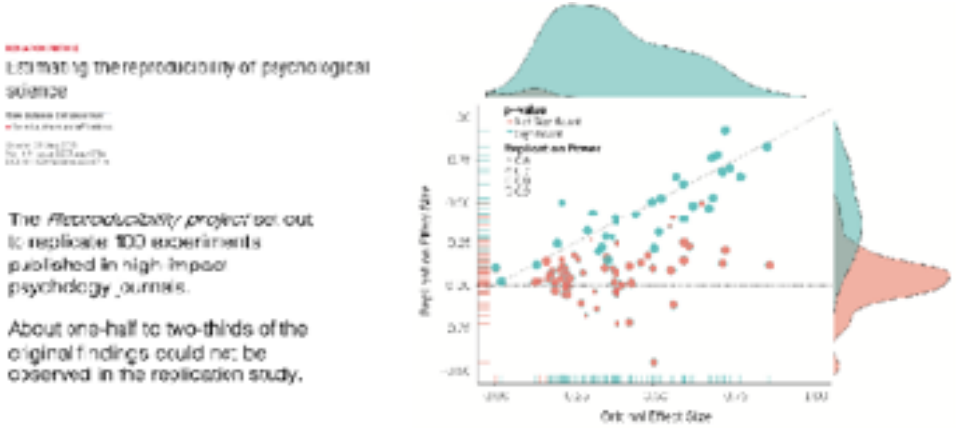


Estimating the reproducibility of psychological science.
Science. **2015**;349: aac4716. doi:10.1126/science.aac4716

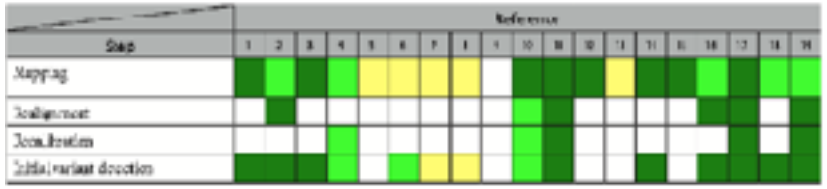
The reproducibility crisis



Repeatability of published microarray gene expression analyses.
Nat Genet. **2009**;41: 149–155. doi:10.1038/ng.295

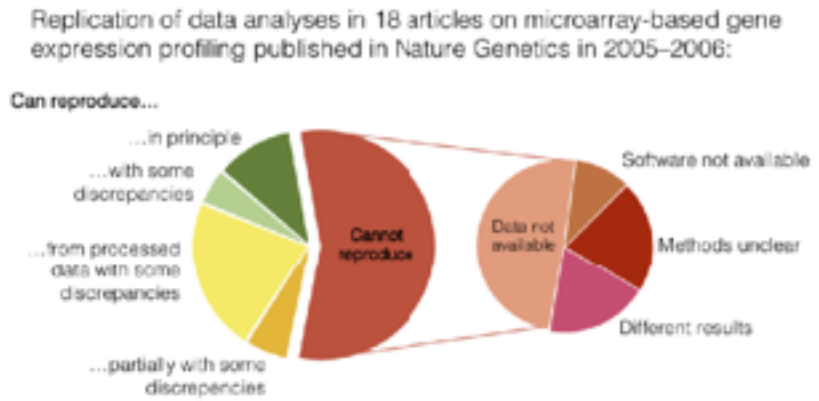


Estimating the reproducibility of psychological science. Science. **2015**;349: aac4716. doi:10.1126/science.aac4716

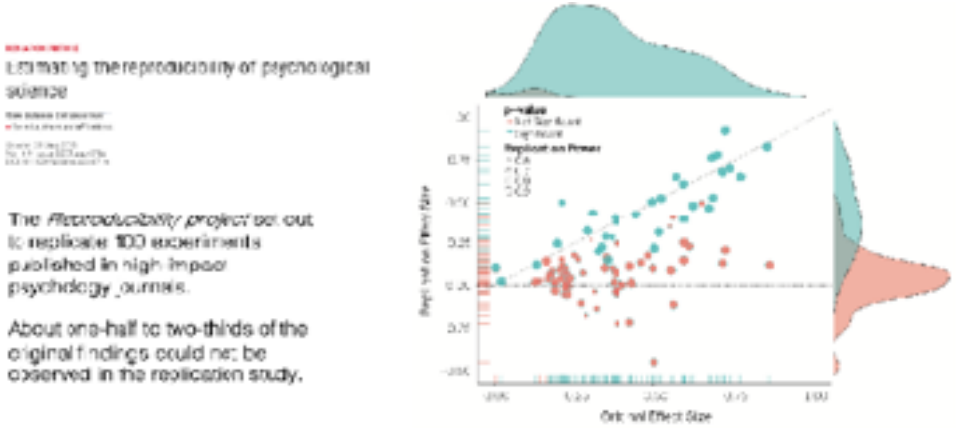


Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. **2012**;13: 667–672.

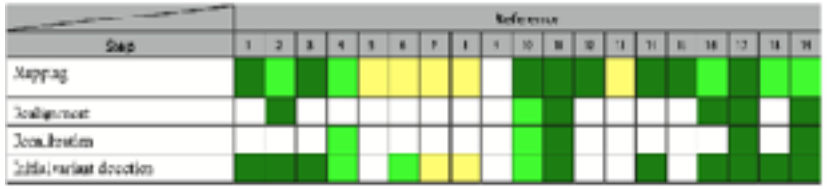
The reproducibility crisis



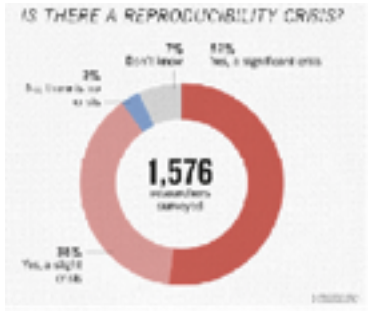
Repeatability of published microarray gene expression analyses.
Nat Genet. **2009**;41: 149–155. doi:10.1038/ng.295



Estimating the reproducibility of psychological science.
Science. **2015**;349: aac4716. doi:10.1126/science.aac4716



Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. **2012**;13: 667–672.



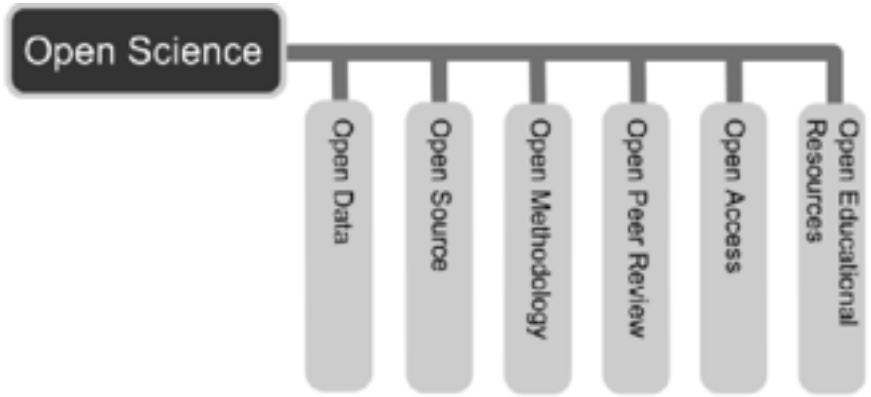
Is there a reproducibility crisis in science?
Nature. **2016**. doi:10.1038/d41586-019-00067-3

Digital Disruption+ Data Deluge + Damage of time + reproducibility crisis

Solution ?

Digital Disruption+ Data Deluge + Damage of time + reproducibility crisis

Solution ?



RESEARCH DATA

What is it ?

https://www.oecd.org/sti/inno/38500813.pdf

RESEARCH DATA

What is it ?

Research data are the **evidence** underlying the **answer to the research question** and can be used to **validate** the results, regardless of their form (i.e. printed, digital or physical).
This may include **quantitative** information or **qualitative** statements collected by researchers in the course of their work through experimentation, observation, modeling, interviewing or other methods, or information derived from existing evidence..

Data may be raw or **primary** (e.g. directly derived from measurements or collections), or **derived** from primary data by analysis or interpretation (e.g. cleaned or extracted from a larger dataset), or derived from existing sources whose rights may be held by others.

https://www.oecd.org/sti/inno/38500813.pdf

What to focus on
for addressing this problem ?

- Data Production & Data Producers
- Data Science and Data Processing
- Data Storage, Management and Distribution

ONCOSPHERE
International Meeting

12-13 janvier 2023

Bordeaux

ONCOSPHERE

Open Science

How to redo **here** and **now** what was done **there before** ?

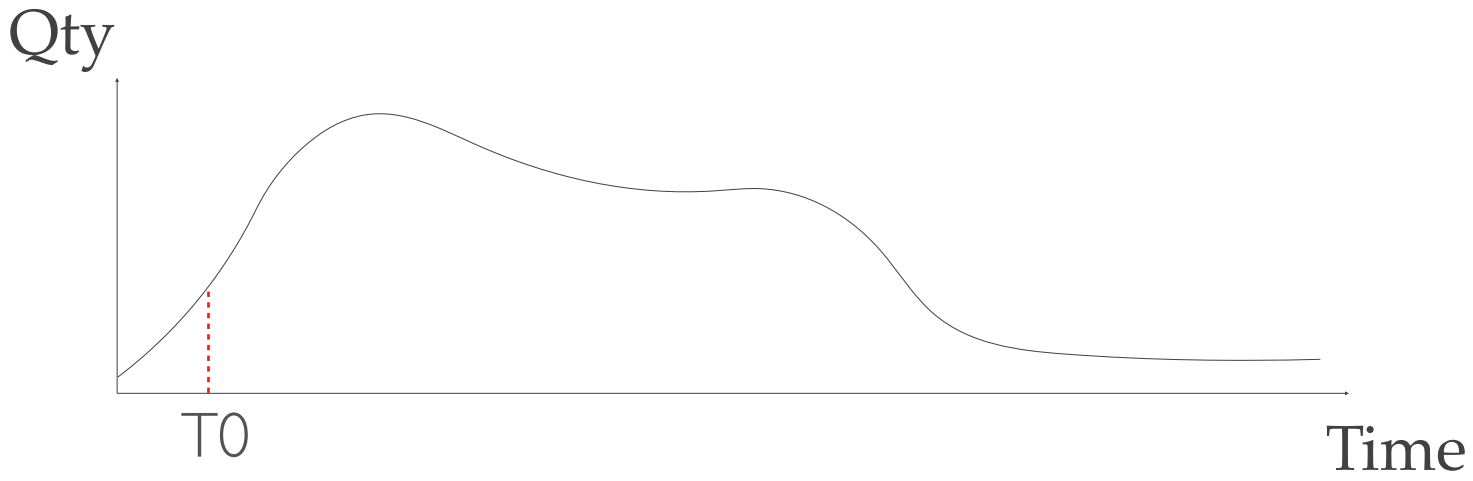
How to proceed **here** and **now** to be able to do it again **later** and **elsewhere** ?

What do I have to do then ?

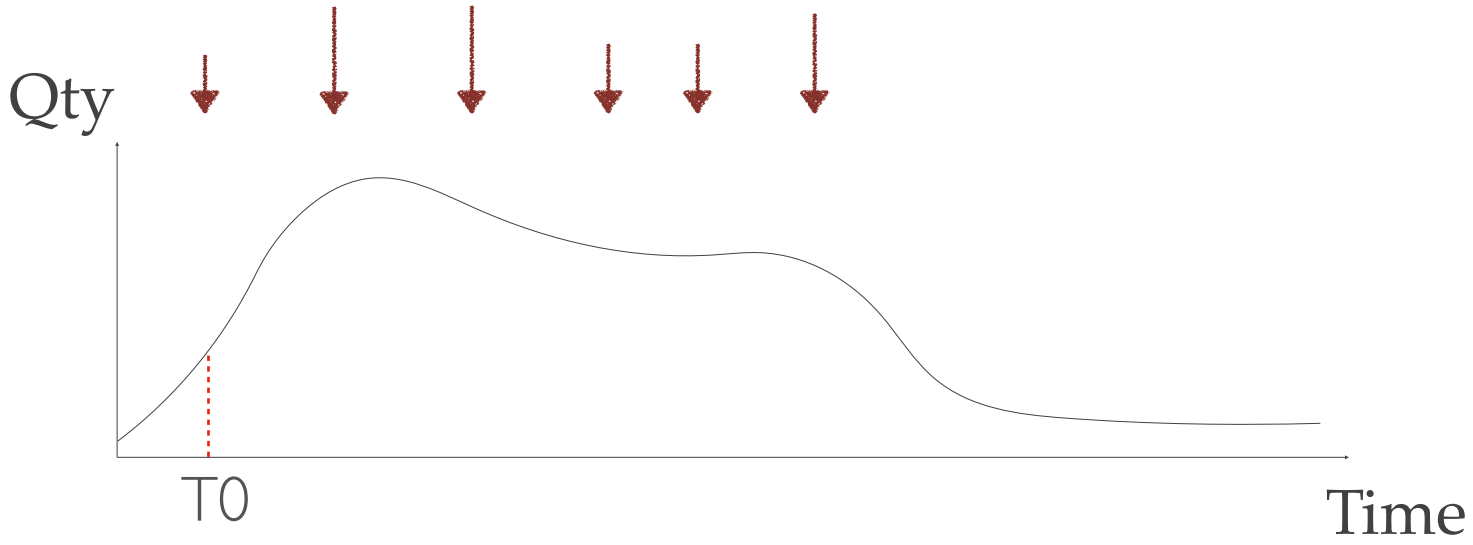
What do I have to do then ?

- Described data so it can be found and understood
- Make data accessible
- Make data reusable
- Make data free

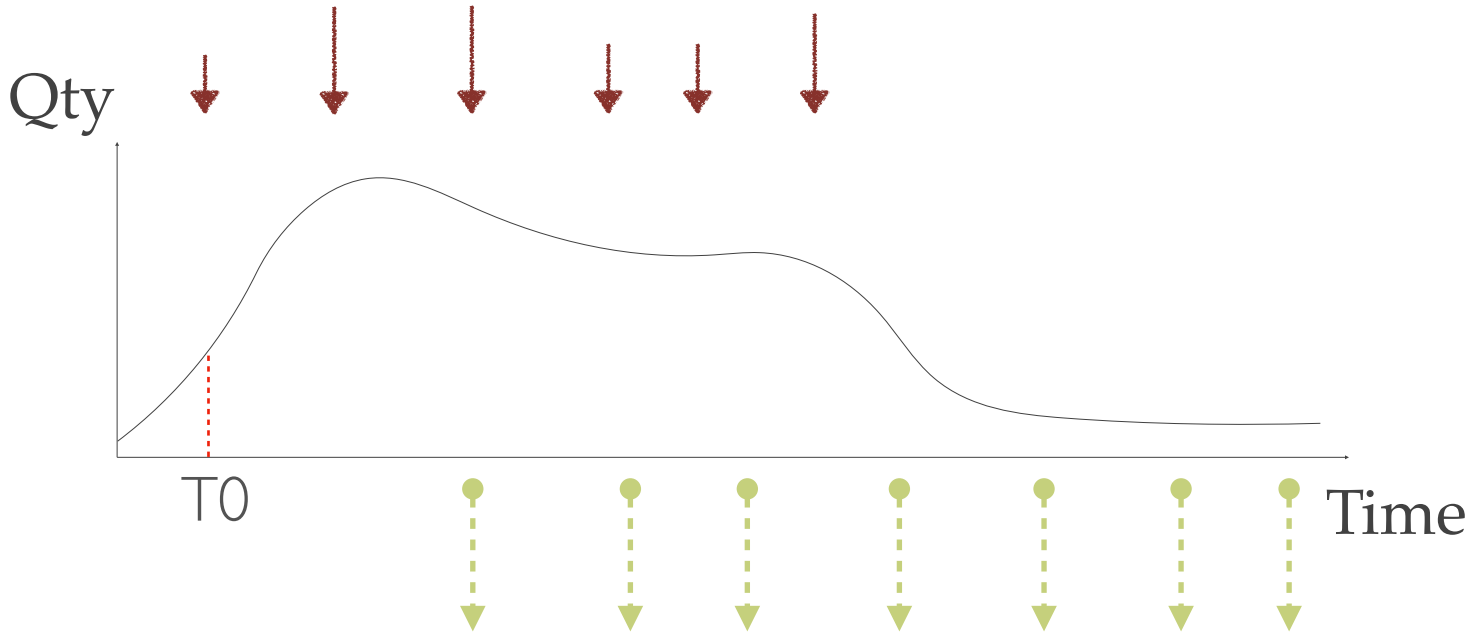
A project over time



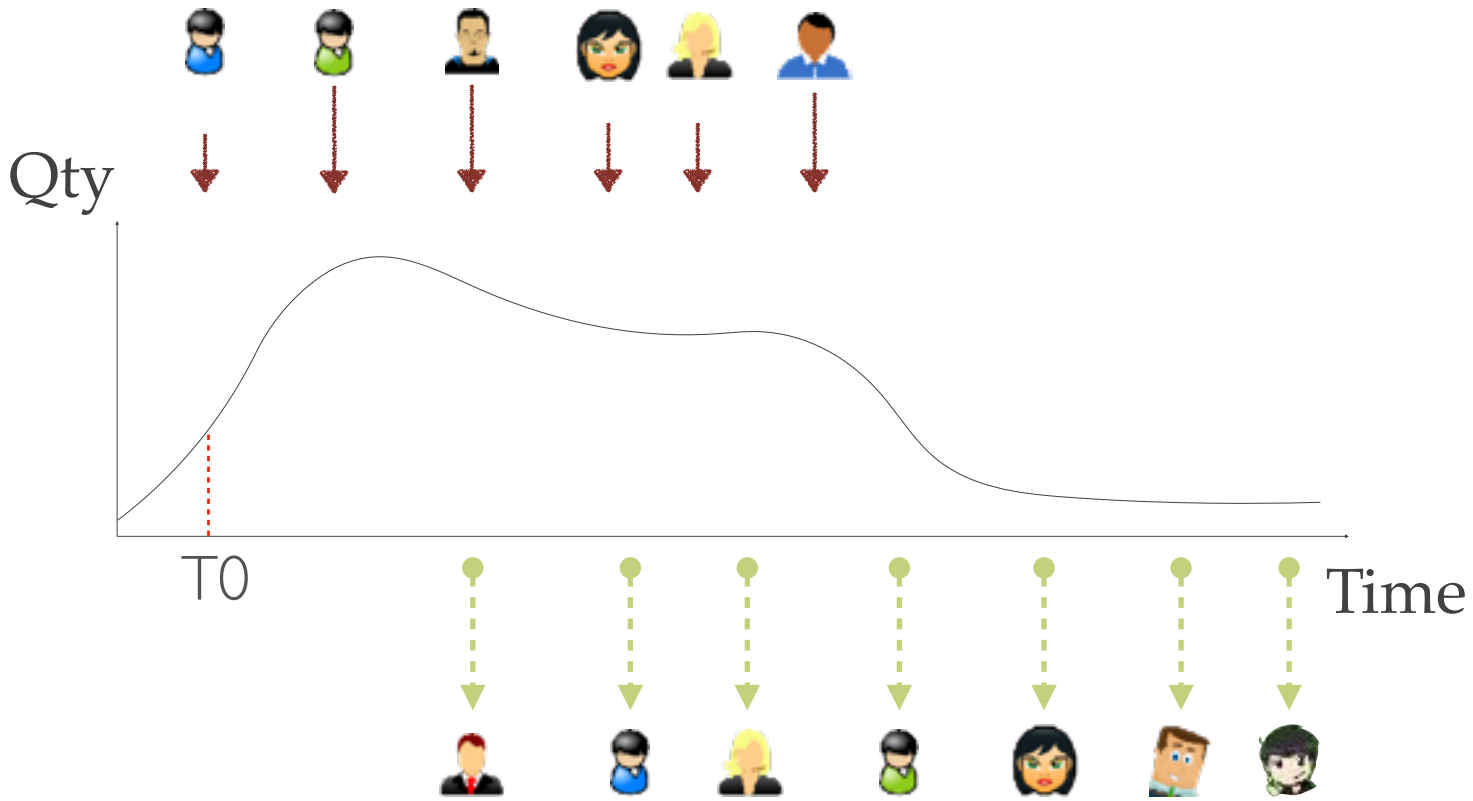
A project over time



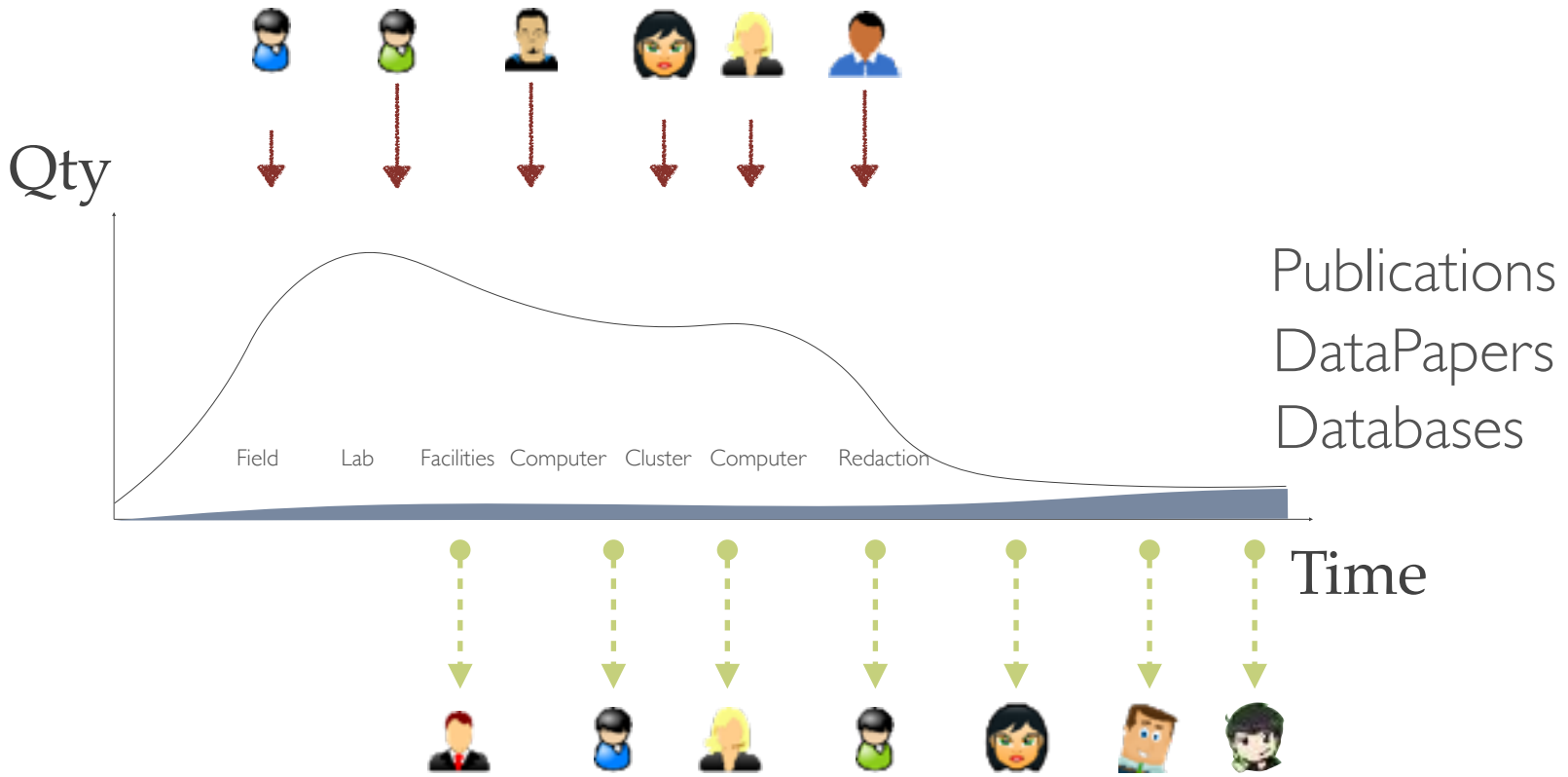
A project over time



A project over time



A project over time



Décrire les données : pour les rendre trouvable et compréhensibles

How to describe your data so that it is identified and found on the web ?

How to explain your data to someone you never met before ?

How to make sure that a prefect stranger will be able to understand your data ?

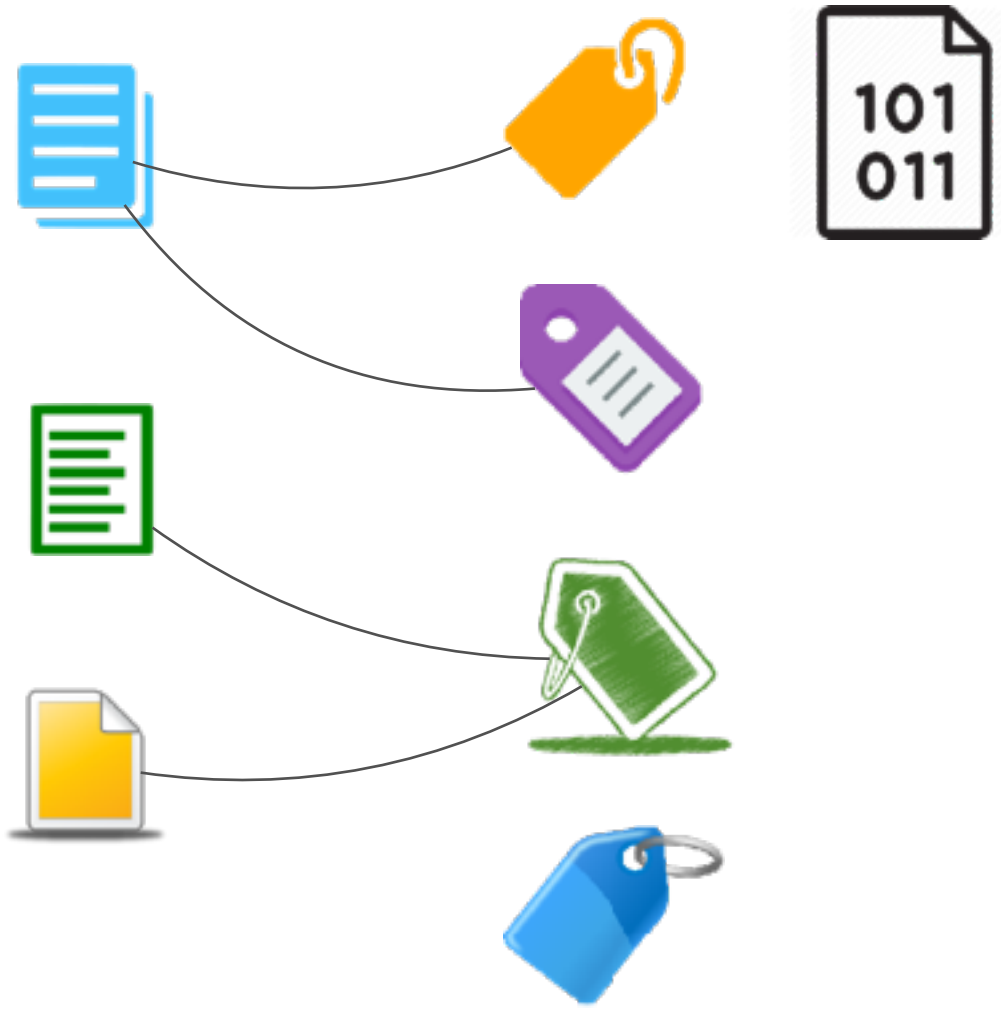
Metadata

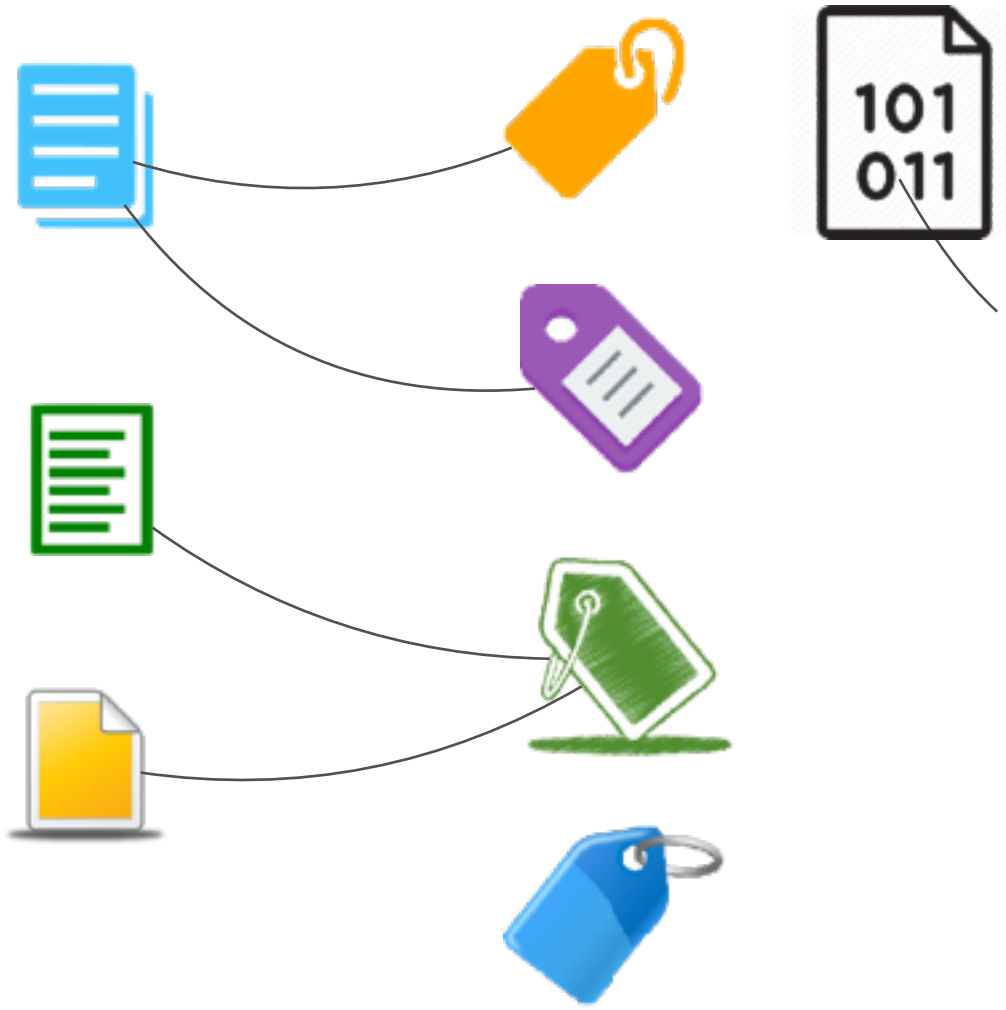
Metadata

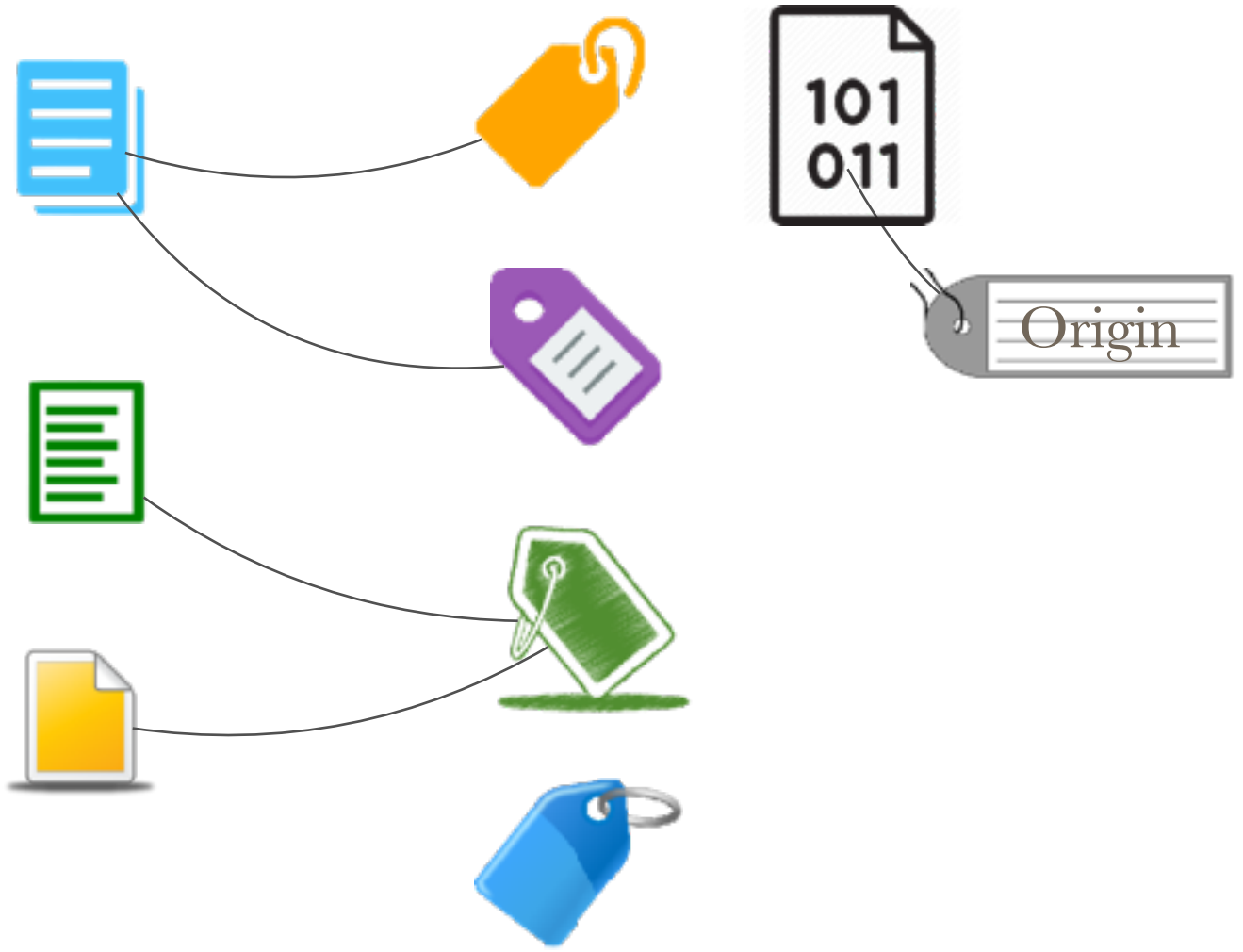
Metadata

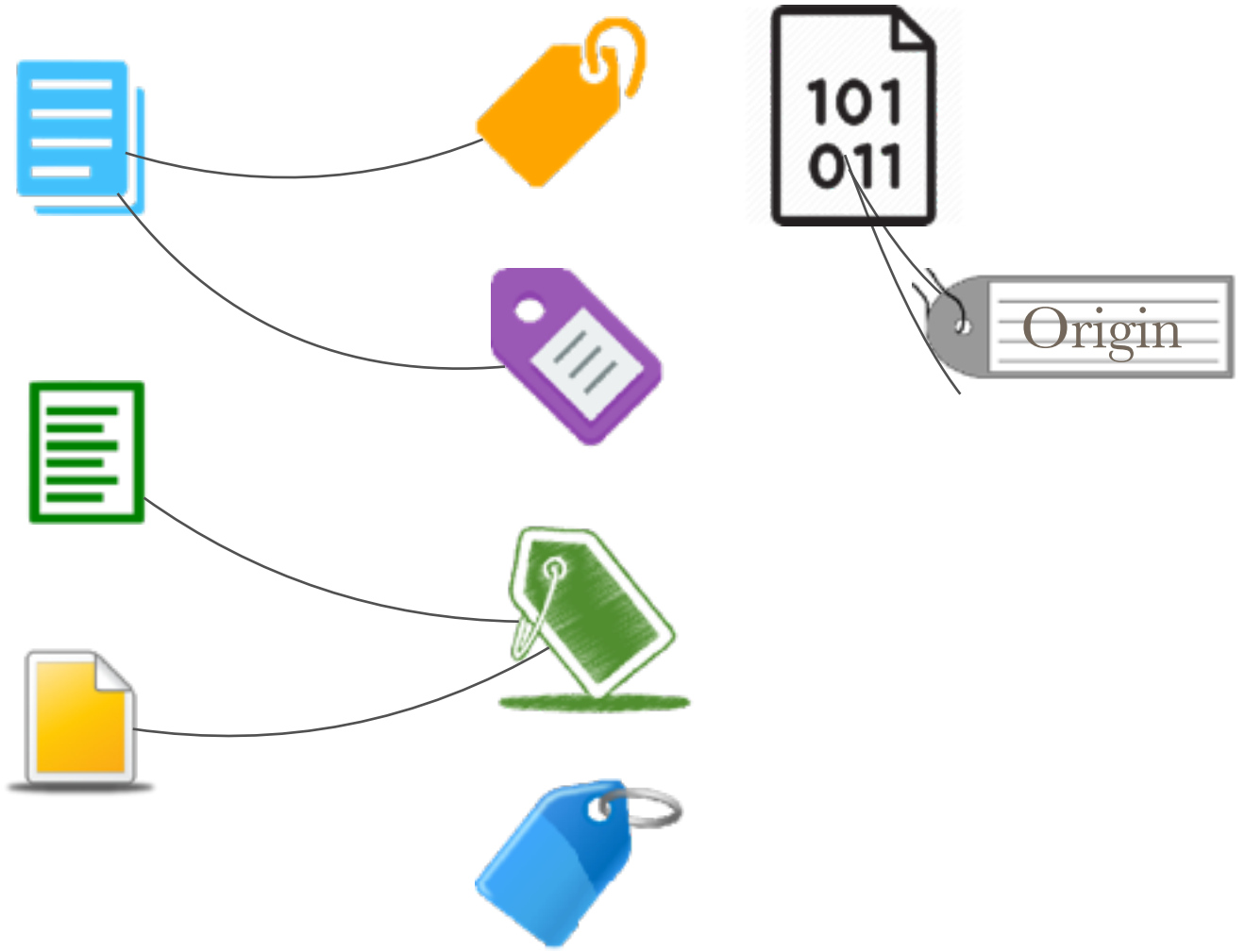

GRAND VIN DE BORDEAUX
2009
Les Fiefs de Lagrange
SAINT-JULIEN

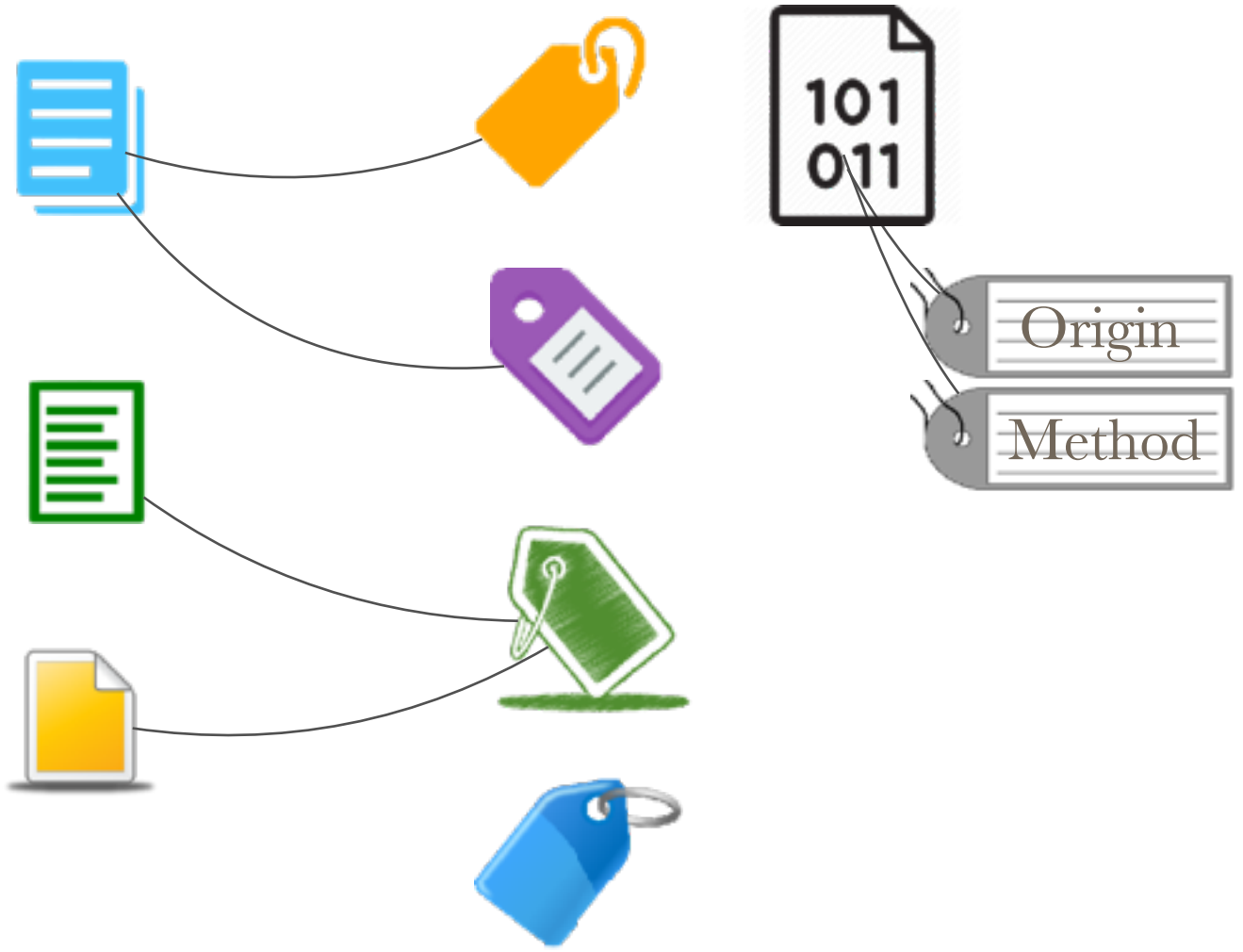

Bonduelle
Petits Pois & Carottes
EXTRA-FINS À L'ÉTUVÉE
Pour 100g
Energie 226 kJ 54 kcal
www.bonduelle.fr

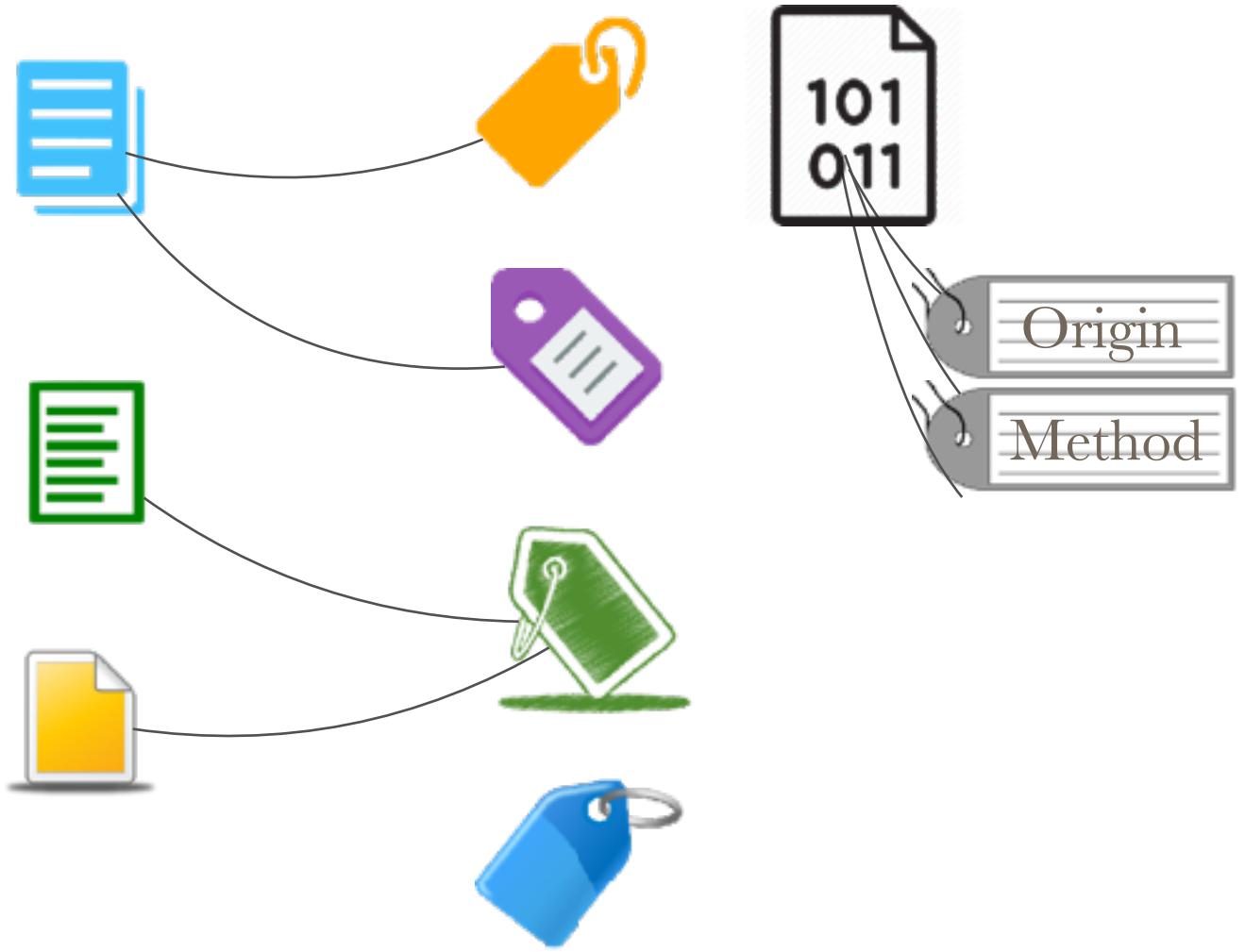

101
011

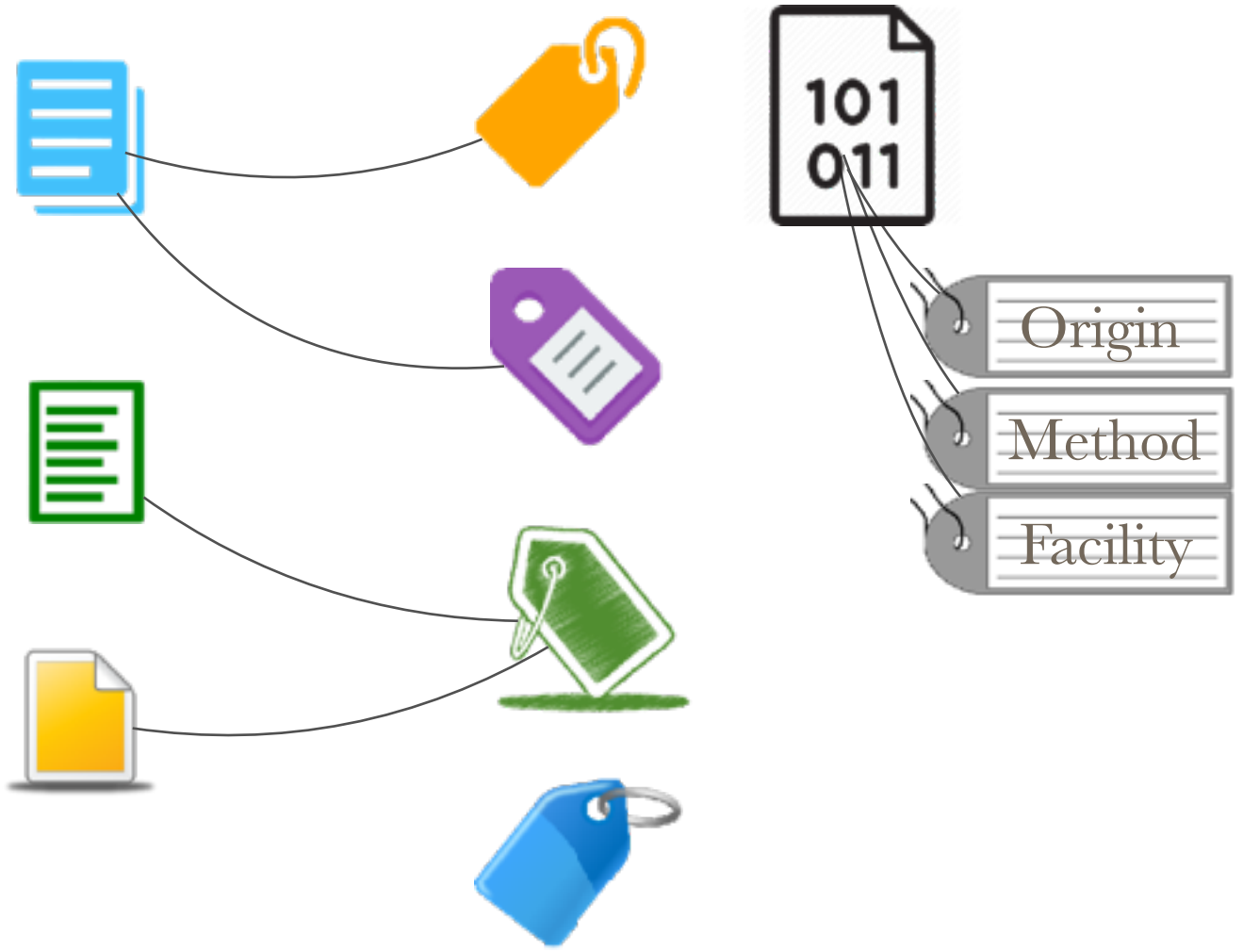

101
011

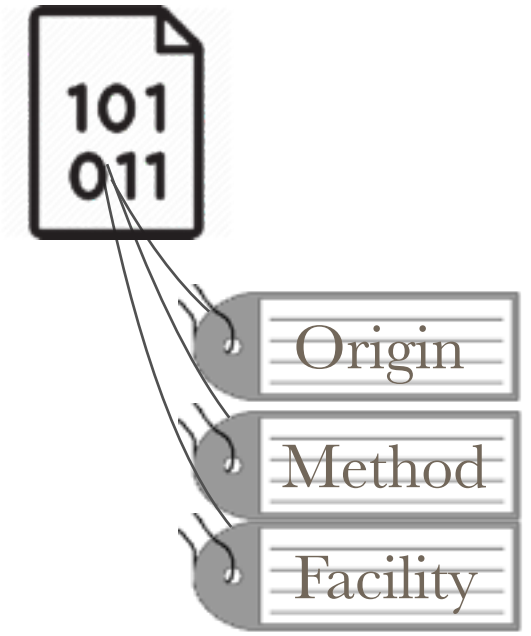

101
011
Origin


101
011
Origin


101
011
Origin
Method


101
011
Origin
Method

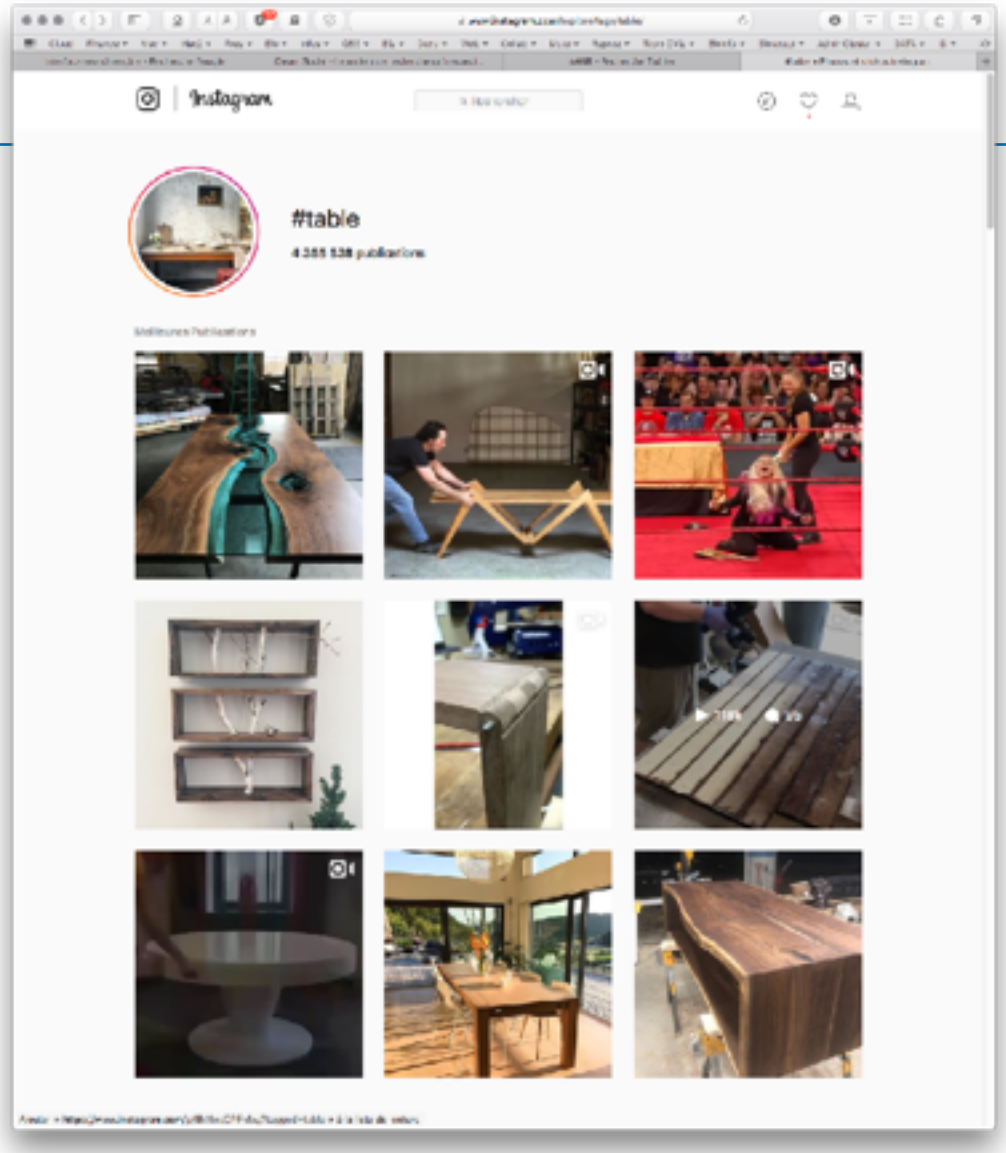

101
011
Origin
Method
Facility



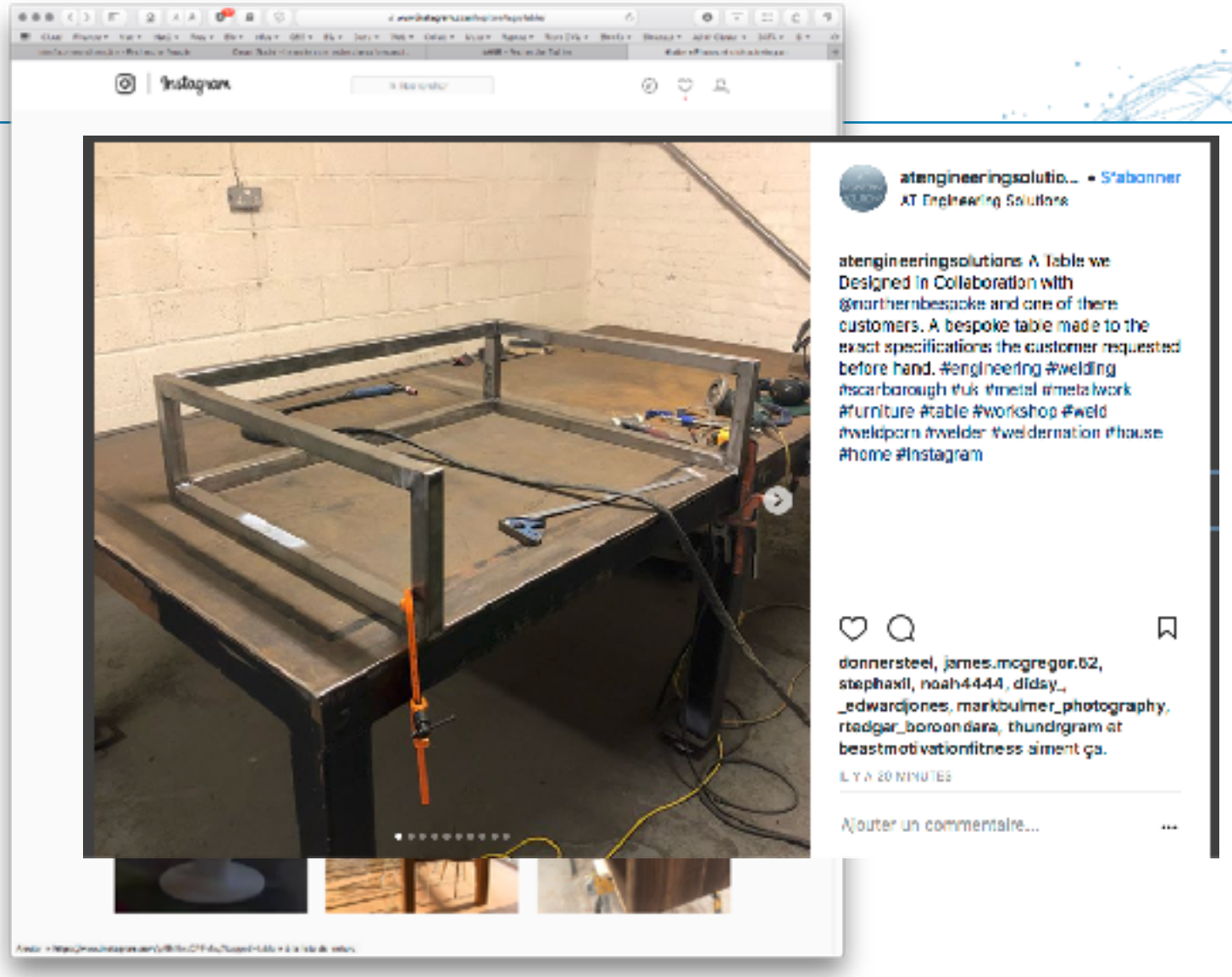
Tags
MetaData

Controlled vocabulary
Defined by the group
Scalable

Make data findable



Make data findable



Controlled Shared Vocabulary

Genus	Species	Sub-species	Group	Name
Oryza	Sativa		japonica	PENTHE BLANC
Oryza	Sativa		japonica	PENTHE NOIR
Oryza	Sativa		indica	ZOGO
Oryza	Glaberrima			GBAI-GBAI
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Musa	acuminata	banksii	wild	Banksii H09

sequençaging Type	Read Length	Equipement	Sequencing location
illumina	1*150	HiSeq3000	Genotoul
illumina	1*150	HiSeq3000	Genotoul

Germplasm	Origin	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

Controlled Shared Vocabulary

List the questions

Genus	Species	Sub-species	Group	Name
Oryza	Sativa		japonica	PENTHE BLANC
Oryza	Sativa		japonica	PENTHE NOIR
Oryza	Sativa		indica	ZOGO
Oryza	Glaberrima			GBAI-GBAI
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Musa	acuminata	banksii	wild	Banksii H09

sequençaging Type	Read Length	Equipement	Sequencing location
illumina	1*150	HiSeq3000	Genotoul
illumina	1*150	HiSeq3000	Genotoul

Germplasm	Origin	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

Controlled Shared Vocabulary

List the questions

List the answers

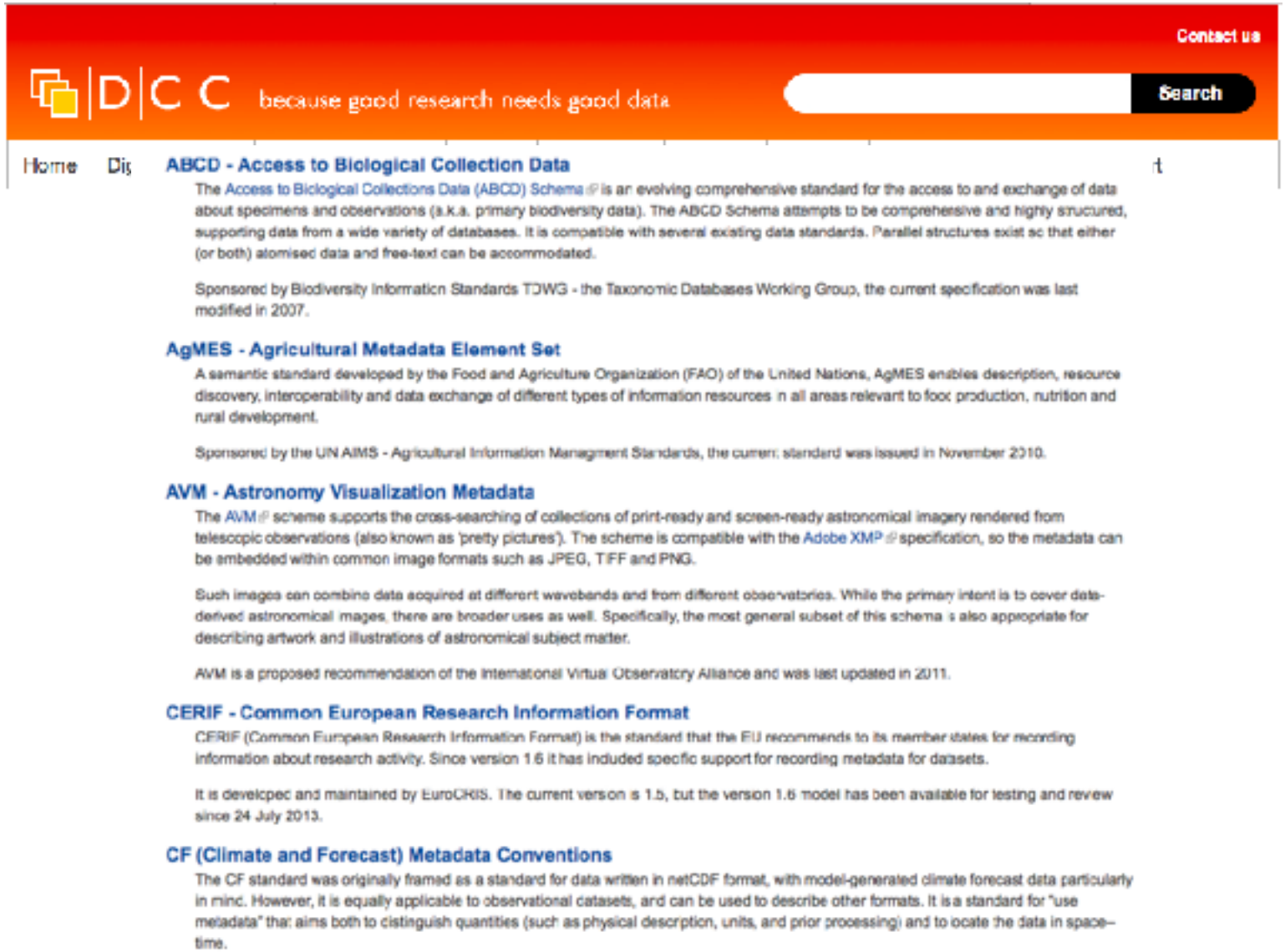
Genus	Species	Sub-species	Group	Name
Oryza	Sativa		japonica	PENTHE BLANC
Oryza	Sativa		japonica	PENTHE NOIR
Oryza	Sativa		indica	ZOGO
Oryza	Glaberrima			GBAI-GBAI
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Musa	acuminata	banksii	wild	Banksii H09

sequençaging Type	Read Length	Equipement	Sequencing location
illumina	1*150	HiSeq3000	Genotoul
illumina	1*150	HiSeq3000	Genotoul

Germplasm	Origin	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

Metadata standards - which one for me ?

Metadata standards - which one for me ?



Contact us

D|C C because good research needs good data

Search

Home Dig t

ABCD - Access to Biological Collection Data

The Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data). The ABCD Schema attempts to be comprehensive and highly structured, supporting data from a wide variety of databases. It is compatible with several existing data standards. Parallel structures exist so that either (or both) atomised data and free-text can be accommodated.

Sponsored by Biodiversity Information Standards TDWG - the Taxonomic Databases Working Group, the current specification was last modified in 2007.

AgMES - Agricultural Metadata Element Set

A semantic standard developed by the Food and Agriculture Organization (FAO) of the United Nations, AgMES enables description, resource discovery, interoperability and data exchange of different types of information resources in all areas relevant to food production, nutrition and rural development.

Sponsored by the UN AIMS - Agricultural Information Management Standards, the current standard was issued in November 2010.

AVM - Astronomy Visualization Metadata

The AVM scheme supports the cross-searching of collections of print-ready and screen-ready astronomical imagery rendered from telescopic observations (also known as 'pretty pictures'). The scheme is compatible with the Adobe XMP specification, so the metadata can be embedded within common image formats such as JPEG, TIFF and PNG.

Such images can combine data acquired at different wavebands and from different observatories. While the primary intent is to cover data-derived astronomical images, there are broader uses as well. Specifically, the most general subset of this schema is also appropriate for describing artwork and illustrations of astronomical subject matter.

AVM is a proposed recommendation of the International Virtual Observatory Alliance and was last updated in 2011.

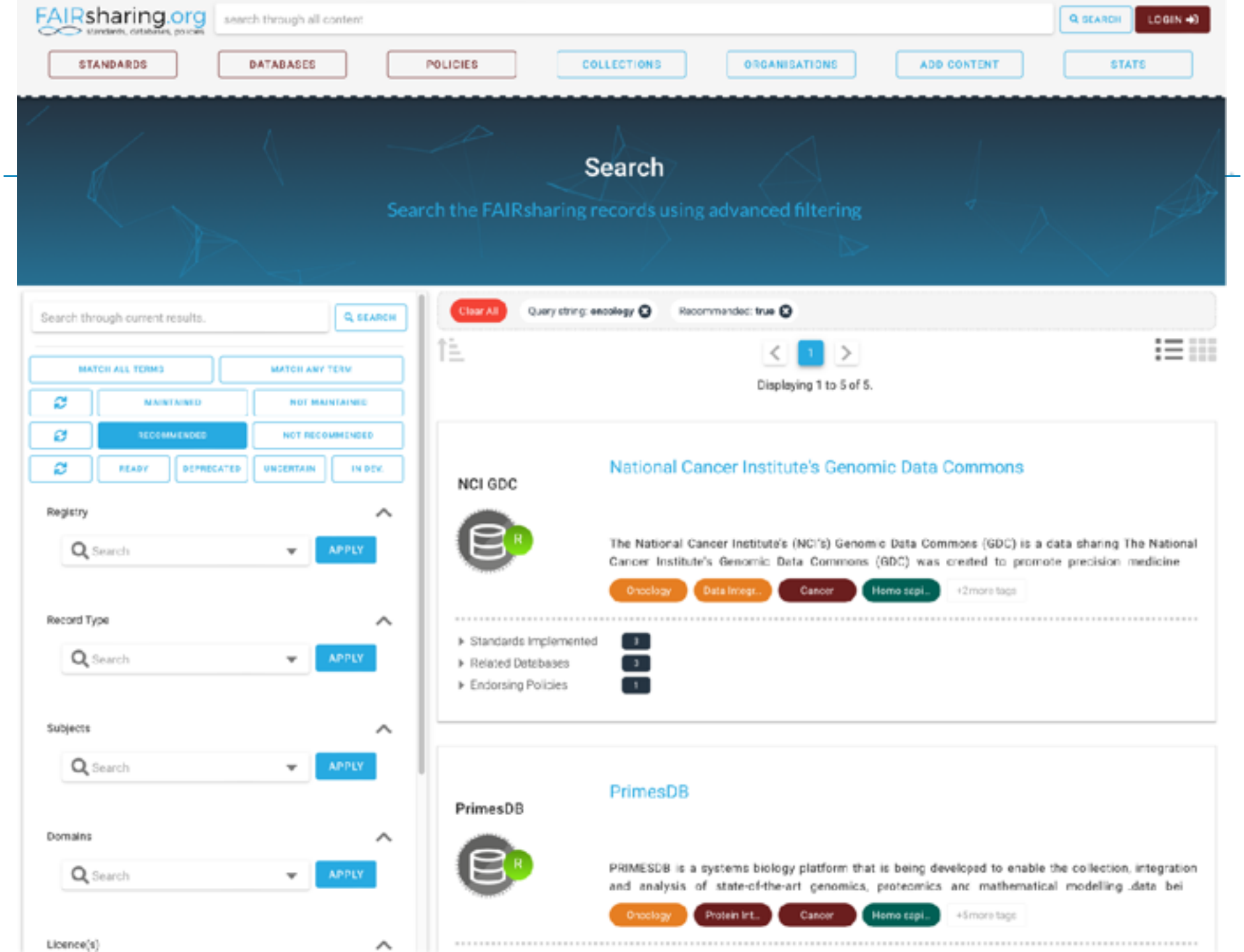
CERIF - Common European Research Information Format

CERIF (Common European Research Information Format) is the standard that the EU recommends to its member states for recording information about research activity. Since version 1.6 it has included specific support for recording metadata for datasets.

It is developed and maintained by EuroCRIS. The current version is 1.5, but the version 1.6 model has been available for testing and review since 24 July 2013.

CF (Climate and Forecast) Metadata Conventions

The CF standard was originally framed as a standard for data written in netCDF format, with model-generated climate forecast data particularly in mind. However, it is equally applicable to observational datasets, and can be used to describe other formats. It is a standard for "use metadata" that aims both to distinguish quantities (such as physical description, units, and prior processing) and to locate the data in space–time.


FAIRsharing.org
search through all content
SEARCH
LOGIN
STANDARDS
DATABASES
POLICIES
COLLECTIONS
ORGANISATIONS
ADD CONTENT
STATS
Search
Search the FAIRsharing records using advanced filtering
Search through current results.
SEARCH
MATCH ALL TERMS
MATCH ANY TERM
MAINTAINED
NOT MAINTAINED
RECOMMENDED
NOT RECOMMENDED
READY
DEPRECATED
UNCERTAIN
IN DEV
Registry
Search
APPLY
Record Type
Search
APPLY
Subjects
Search
APPLY
Domains
Search
APPLY
Licence(s)
Clear All
Query string: oncology
Recommended: true
1
Displaying 1 to 5 of 5.
NCI GDC
National Cancer Institute's Genomic Data Commons
The National Cancer Institute's (NCI's) Genomic Data Commons (GDC) is a data sharing The National Cancer Institute's Genomic Data Commons (GDC) was created to promote precision medicine
Oncology
Data Integr...
Cancer
Homo sapi...
+2 more tags
Standards Implemented
3
Related Databases
3
Endorsing Policies
1
PrimesDB
PrimesDB
PRIMESDB is a systems biology platform that is being developed to enable the collection, integration and analysis of state-of-the-art genomics, proteomics and mathematical modelling .data bei
Oncology
Protein Int...
Cancer
Homo sapi...
+5 more tags

GENERAL INFORMATION

National Cancer Institute's Proteomic Data Commons (NCI PDC)

10.25504/FAIRsharing.kXzfjt

Type	Repository
Registry	Database
Description	The NCI's Proteomic Data Commons has been created to make cancer-related proteomic datasets easily accessible to the public, and to facilitate multi-omic integration in support of precision medicine through interoperability with other resources. The PDC is one of several repositories within the NCI Cancer Research Data Commons (CRDC), a secure cloud based infrastructure featuring diverse data sets and innovative analytic tools – all designed to advance data-driven scientific discovery. The CRDC enables researchers to link proteomic data with other data sets (e.g., genomic and imaging data) and to submit, collect, analyze, store, and share data throughout the cancer data ecosystem.
Homepage	https://pdc.cancer.gov/
Year of Creation	2018
Maintainers	thangudu
Countries developing this resource	United States
Subjects	Oncology, Proteomics, Proteogenomics
Domains	Cancer
Taxonomic Range	Homo sapiens
User Defined Tags	None

VIEW RELATION GRAPH

How to cite this record

FAIRsharing.org: NCI PDC; National Cancer Institute's Proteomic Data Commons, **DOI:** 10.25504/FAIRsharing.kXzfjt, **Last Edited:** Tuesday, August 16th 2022, 10:03, **Last Editor:** delphinedauga, **Last Accessed:** Saturday, January 7th 2023, 21:28

Record ID 2989 | **Record created at** Monday, May 11th 2020, 16:30 | **Record updated at** Tuesday, August 16th 2022, 10:03

Pour rendre les données accessibles, il faut

Keep your data safe

Back Up, Access Control

Keep your data accessible

Run a website

All the time

24/7

For Ever …

Simple Answer : **Repositories**

What is a data repository?

A data repository is a storage space for researchers to deposit data sets associated with their research. And if you're an author seeking to comply with a journal **data sharing policy**, you'll need to identify a suitable repository for your data.

Data Repositories for Research

- **Institutional** Dataverse
- **Europe** Zenodo, B2Share
- **Global** Figshare, Dryad
- **Editors** Oxford Univ Press (GigaDB) ; Ubiquity Press (Dataverse)
- **Specialized**
  - GBIF (Global Biodiversity Information Facility)
  - KNB (Knowledge Network for biocomplexity), EDI (Environmental Data Initiative)
  - Pangaea, SEANOE
  - Movebank, WormBase, ViPR, MycoBank, ComBase, FLOW
  - GenBank, Barcode of Life Data Systems, UniProt, Intact
  - Dataverse, ICPSR, DataFirst, Quetelet, beQuali

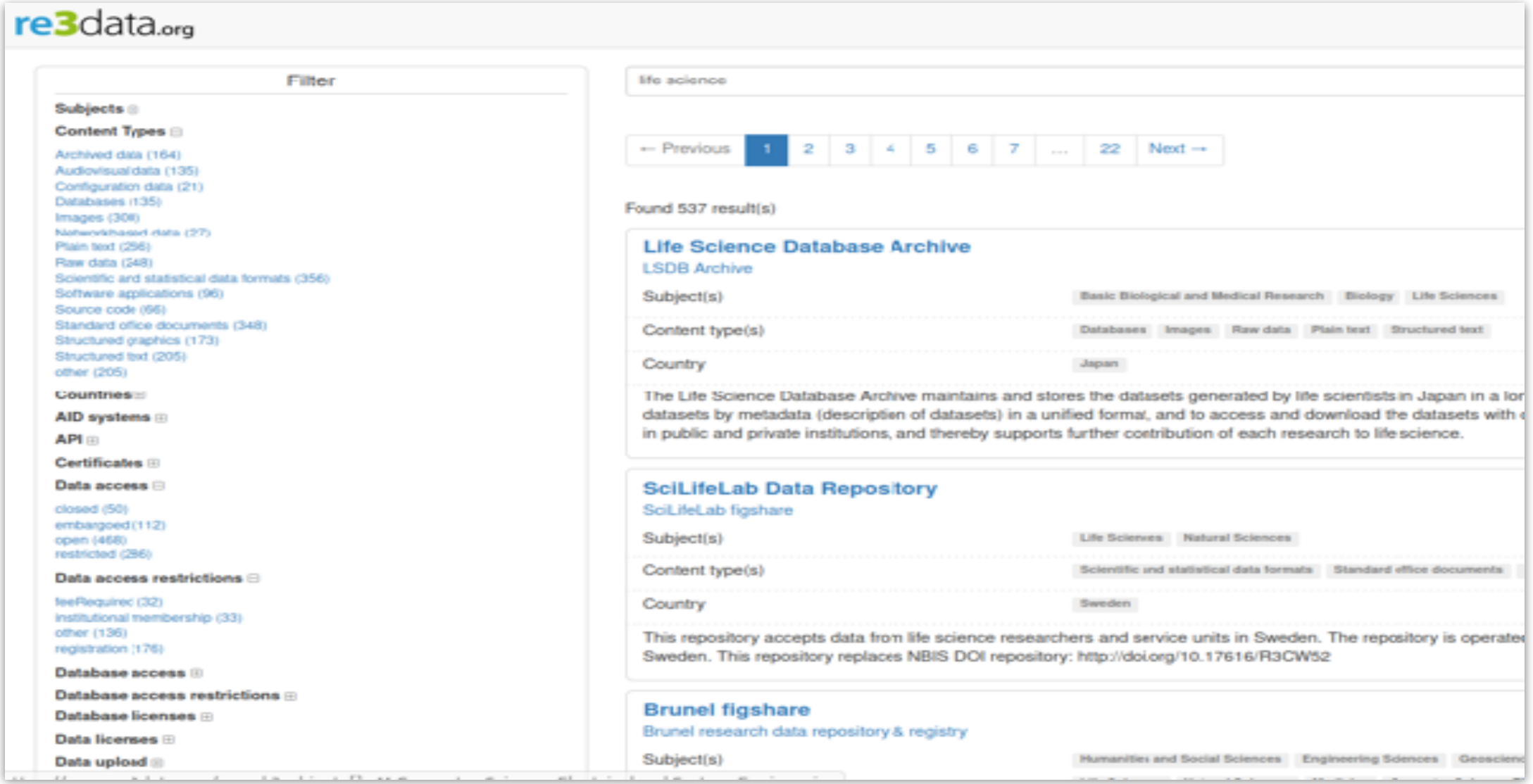
- Recherche Data Gouv, the federated national research data platform

Repository Services

- Secured storage for your data
- Easy access (web browser)
- Access control
- Harvested by catalogs robots
- Available 24/7

- And they ask for metadata … you just added tou your data !

How to find repositories



Filter

Subjects
Content Types
Countries
AID systems
API
Certificates
Data access

closed (3)
embargoed (5)
open (25)
restricted (19)

Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards

DDI - Data Documentation Initiative (1)
DataCite Metadata Schema (1)
Dublin Core (2)
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (1)
ISA-Tab (1)
OAI-ORE - Open Archives Initiative Object Reuse and Exchange (1)
Repository-Developed Metadata Schemas (2)

PID systems
Provider types
Quality management

oncology

Search

Toogle short help

← Previous 1 2 Next →

Sort by

Found 32 result(s)

Project Data Sphere

PDS

Subject(s): Radiation Oncology and Radiobiology | Medicine | Public Health, Health Services Research, Social Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Networkbased data | Software applications | Standard office documents | Scientific and statistical data formats | Raw data

Country: United States

Project Data Sphere, LLC, operates a free digital library-laboratory where the research community can broadly share, integrate and analyze historical, de-identified, patient-level data from academic and industry cancer Phase II-III clinical trials. These patient-level datasets are available through the Project Data Sphere platform to researchers affiliated with life science companies, hospitals and institutions, as well as independent researchers, at no cost and without requiring a research proposal.

Netherlands Cancer Registry

NCR

Subject(s): Epidemiology, Medical Biometry, Medical Informatics | Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Raw data | Scientific and statistical data formats | Standard office documents

Country: Netherlands

The Netherlands Cancer Registry is the national registration since 1989, providing statistics on cancer in the Netherlands. The registry is maintained by the Netherlands Comprehensive Cancer Organisation (IKNL). Data on incidence, prevalence, survival, mortality can be viewed in NCR data & figures on the IKNL website.

National Cancer Data Base

Filter

Subjects
Content Types
Countries
AID systems
API
Certificates
Data access

closed (3)
embargoed (5)
open (25)
restricted (19)

Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards

DDI - Data Documentation Initiative (1)
DataCite Metadata Schema (1)
Dublin Core (2)
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (1)
ISA-Tab (1)
OAI-ORE - Open Archives Initiative Object Reuse and Exchange (1)
Repository-Developed Metadata Schemas (2)

PID systems
Provider types
Quality management

oncology

Search

Toogle short help

← Previous 1 2 Next →

Sort by

Found 32 result(s)

Project Data Sphere

PDS

Subject(s): Radiation Oncology and Radiobiology | Medicine | Public Health, Health Services Research, Social Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Networkbased data | Software applications | Standard office documents | Scientific and statistical data formats | Raw data

Country: United States

Project Data Sphere, LLC, operates a free digital library-laboratory where the research community can broadly share, integrate and analyze historical, de-identified, patient-level data from academic and industry cancer Phase II-III clinical trials. These patient-level datasets are available through the Project Data Sphere platform to researchers affiliated with life science companies, hospitals and institutions, as well as independent researchers, at no cost and without requiring a research proposal.

Netherlands Cancer Registry

NCR

Subject(s): Epidemiology, Medical Biometry, Medical Informatics | Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Raw data | Scientific and statistical data formats | Standard office documents

Country: Netherlands

The Netherlands Cancer Registry is the national registration since 1989, providing statistics on cancer in the Netherlands. The registry is maintained by the Netherlands Comprehensive Cancer Organisation (IKNL). Data on incidence, prevalence, survival, mortality can be viewed in NCR data & figures on the IKNL website.

National Cancer Data Base

Filter

Subjects
Content Types
Countries
AID systems
API
Certificates
Data access

closed (3)
embargoed (5)
open (25)
restricted (19)

Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards

DDI - Data Documentation Initiative (1)
DataCite Metadata Schema (1)
Dublin Core (2)
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (1)
ISA-Tab (1)
OAI-ORE - Open Archives Initiative Object Reuse and Exchange (1)
Repository-Developed Metadata Schemas (2)

PID systems
Provider types
Quality management

oncology

Search

Toogle short help

← Previous 1 2 Next →

Sort by

Found 32 result(s)

Project Data Sphere

PDS

Subject(s): Radiation Oncology and Radiobiology | Medicine | Public Health, Health Services Research, Social Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Networkbased data | Software applications | Standard office documents | Scientific and statistical data formats | Raw data

Country: United States

Project Data Sphere, LLC, operates a free digital library-laboratory where the research community can broadly share, integrate and analyze historical, de-identified, patient-level data from academic and industry cancer Phase II-III clinical trials. These patient-level datasets are available through the Project Data Sphere platform to researchers affiliated with life science companies, hospitals and institutions, as well as independent researchers, at no cost and without requiring a research proposal.

Netherlands Cancer Registry

NCR

Subject(s): Epidemiology, Medical Biometry, Medical Informatics | Medicine | Hematology, Oncology, Transfusion Medicine | Medicine | Life Sciences

Content type(s): Raw data | Scientific and statistical data formats | Standard office documents

Country: Netherlands

The Netherlands Cancer Registry is the national registration since 1989, providing statistics on cancer in the Netherlands. The registry is maintained by the Netherlands Comprehensive Cancer Organisation (IKNL). Data on incidence, prevalence, survival, mortality can be viewed in NCR data & figures on the IKNL website.

National Cancer Data Base

Repository details

CancerData.org

General | Institutions | Terms | Standards

Name of repository	**CancerData.org**
Additional name(s)	Sharing data for cancer research
Repository URL	https://www.cancerdata.org/
Subject(s)	Basic Biological and Medical Research; Medicine; Biology; Life Sciences
Description	The CancerData site is an effort of the Medical Informatics and Knowledge Engineering team (MIKE for short) of Maastro Clinic, Maastricht, The Netherlands. Our activities in the field of medical image analysis and data modelling are visible in a number of projects we are running. CancerData is offering several datasets. They are grouped in collections and can be public or private. You can search for public datasets in the NBIA (National Biomedical Imaging Archive) image archives without logging in.
Contact	erik.roelofs@maastro.nl
Content type(s)	Standard office documents; Databases; Images; Structured graphics; Scientific and statistical data formats; Raw data; Plain text; Archived data; other
Keyword(s)	Computer Assisted Theragnostics CAT; cancer; image; medicine; radiotherapy; treatment; tumor
Persistent identifier(s) of the repository	FAIRsharing_doi:10.25504/FAIRsharing.s2txbp
Repository size	522 datasets
Repository type(s)	disciplinary
Mission statement for designated community	http://www.maastro.nl/en/1/77/strategienota.aspx
Research data repository language(s)	English

Repository details

CancerData.org

General | Institutions | Terms | Standards

Database access

Type of access to research data repository	open

Database licenses (1)

Database License	CC

Data access (3)

Type of access to data	open
Type of access to data	restricted
Data access restriction type(s)	registration
Type of access to data	closed

Data licenses (1)

DataLicense	CC

Data upload (1)

Type of data upload	restricted
Data upload restriction type(s)	registration

Repository details

CancerData.org

General | Institutions | Terms | Standards

Persistent identifier system(s)	DOI
Name of the repository software	other
Versioning	yes
Enhanced Publication	yes
Quality management	yes

Application programming interfaces (3)

API type	other
API type	OAI-PMH
API type	FTP

Remarks

Remarks	The MIKE team encourages the use of Free and Open Source Software (FOSS). The CancerData site is build with a DRUPAL (https://www.drupal.org/) web portal, caBIG (http://cbiit.nci.nih.gov/ncip) data services and running on Debian (http://www.debian.org/) servers.
Entry date	2014-06-25
Last update	2021-10-11

Back to search | Submit a change request | Get a badge

Should we use the publisher’s repositories 😒 ?

Research Data Policies

Data policy types

Data availability statements

Data policy FAQs

Journal policies & services

Recommended repositories list

Research Data Helpdesk

Research Data Support

Recommended Repositories

In general, data should be submitted to discipline-specific, community-recognised repository where possible, or to generalist repositories if no suitable community resource is available.

This list is derived from *Scientific Data*'s recommended repository list. The list is also available in figshare for free reuse, with attribution, by others. *Scientific Data* is an open access data journal published by Springer Nature.

Authors should consult journal information for authors in case of more specific repository recommendations. Some Springer Nature journals, such as *Human Genome Variation*, maintain their own research data repositories.

If an author's preferred repository is not listed below we encourage repository managers to investigate listing their repositories with re3data.org and/or FAIRsharing.org in the first instance and, for those that meet the criteria, applying for listing with *Scientific Data*.

For more information browse our repository FAQs.

Find the best repository for your data

Biological sciences
Chemistry and Chemical biology
Earth, Environmental and Space sciences
Generalist repositories
Health sciences
Materials science
Physics
Social science
Other repositories

Earth, Environmental & Oceanographic Data

Data Repository	How articles and data are linked (learn more)	More information
Arctic Data Center (NSF Arctic Data Center)	Authors should include data DOIs in their manuscript, following Publication Guidelines.	ADC homepage Publication Guidelines
BGS Geochem	Authors should specify BGS Geochem sample numbers, e.g. GA000 MAT1001.	BGS Geochem homepage
Carbon Dioxide Information Analysis Center (CDIAC)	Authors should include data DOIs in their manuscript	CDIAC homepage Publication Guidelines Example article
Natural Environment Research Council (NERC), including BADC, BGDC, CEDA and NCEDC	Authors should include data DOIs in their manuscript.	NERC Data Centres Example article
System for Earth Sample Registration (SESAR), registry for International Geo Sample Numbers (IGSN)	Authors should specify IGSN numbers, e.g. IGSN: 110000XXX.	SESAR homepage Registering samples Example article

Engineering & Materials Standards

Data Repository	How articles and data are linked	More information
ASTM (formerly American Society for Testing and Materials)	Authors should specify ASTM Data Files, e.g. ASTM: 003.	ASTM standards homepage Example article

Genes & Gene Expression

Data Repository	How articles and data are linked	More information
Allele Frequency Net Database (AFND)	Authors should specify AFND accession numbers, e.g. AFND: APINDIC1243.	AFND homepage Submitting data
ArrayExpress	Authors should specify ArrayExpress accession numbers, e.g. ArrayExpress: E-MEXP-1783.	ArrayExpress homepage Submitting data Example article
GenBank	Authors should specify GenBank accession numbers, e.g. GenBank: AA123456. ScienceDirect displays and visualizes supporting information made available from and linking to the repository.	GenBank homepage Submitting data
Gene Expression Omnibus (GEO)	Authors should specify GEO accession numbers, e.g. GEO: GSE12345, GPL1234, GSM12345. ScienceDirect displays supporting information made available from and linking to the repository.	GEO homepage Submitting data Example article
Genome Sequence Archive	Authors should specify GSA identifiers, e.g. GSA: CRA000123.	GSA homepage

Interdisciplinary

Data Repository	How articles and data are linked	More information
figshare	Authors should include data DOIs in their manuscript.	figshare homepage Publishing data Publishing tools
Dataverse	Dataverse hosts institutional Dataverse repositories as well as numerous individual Dataverses. Users store their own data and link to it from their publication. Authors should include the dataset DOI in the article, e.g. doi:10.7910/DVN/26888	Harvard Dataverse page Publishing data Example article
Mendeley Data	Mendeley Data hosts research data of any type in any discipline. Authors should include the DOI for their deposited data in the article.	Mendeley Data homepage Publishing data Example article

Life Sciences

Data Repository	How articles and data are linked	More information
Dryad	Authors should specify Dryad accession numbers, e.g. Dryad: doi:10.5061/dryad.123	Publishing data
FlyBase	Authors should specify FlyBase accession numbers, e.g. FlyBase: FBgn0000490.	FlyBase homepage Publishing data
MGI Database	Authors should specify MGI accession numbers, e.g. MGI Database: 4838.	MGI Database homepage Publishing data
PANGAEA	Authors should specify PANGAEA accession numbers, e.g. PANGAEA: 788537.	PANGAEA homepage Publishing data
TreeBASE	Authors should specify TreeBASE accession numbers, e.g. TreeBASE: TB2:S1234.	TreeBASE homepage Publishing data
UK PubMed Central	Authors should specify UKPMC accession numbers, e.g. UKPMC: PMC1234.	UKPMC homepage Publishing data

Materials Science

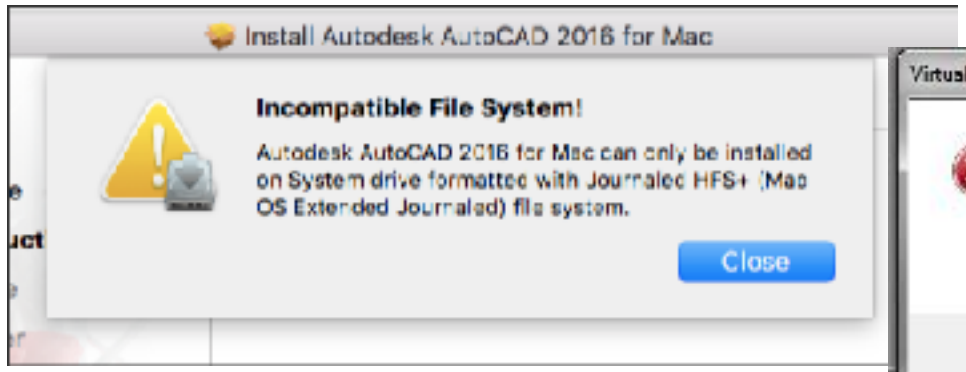
Data Repository	How articles and data are linked	More information
Protein Data Bank (PDB)	Authors should specify Data DOI in their manuscript.	PDB homepage Publishing data

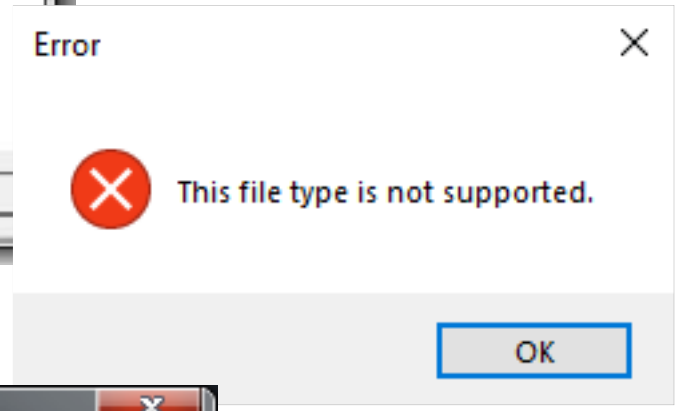
Model Organisms

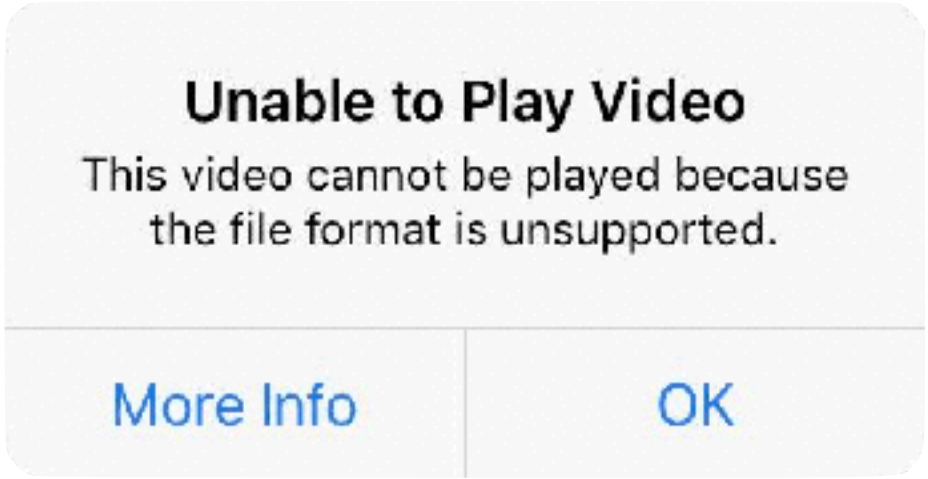
Data Repository	How articles and data are linked	More information

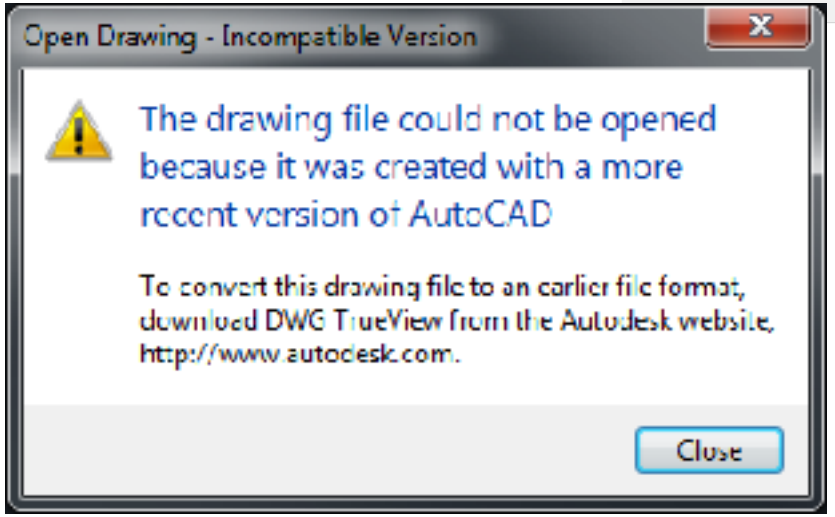
Make your data reusable

Make your data reusable










scrumblr by aliasaria
SAM
cram
html
tsv
csv
doc, docx
ppt, pptx
.tex
SVG
BAM
BAI
fasta
csv/tsv
TXT
png, jpeg, tiff
markdown / md
bed
Bigwig
pour les anciens fasta
mov, mp3, mp4
fastq
pdf
CRAM
mkv
FASTA
VCF
BCF
xml
PDB
mol, mol2
YAML
gbff
rdf
json
R
RData
gif
GFF
PhyloXML
tree
java
.sh
ipynb
nf

File format : think ahead

Is a specific software required to process the format you use ?

Do they work online or after installation on a computer?

Do they work with a particular operating system (Windows, Mac, Linux)?

Or a version of it (XP, Win 7 …)

Are they related to a type of computer or a particular instrument (e.g. microscope)?

Are they free or paid? Who pays ?

If they no longer existed or if you no longer have access to them, could you continue to work?

Is the software publisher (or the community) healthy?

Will the software still be available in 20 years and will it still be able to interpret the file correctly?

What do you suggest to guarantee the sustainability of interpretation to your data?

What are the consequences for you ?

Or the next generation ?

Recommended formats

Type	Formats conseillés	Formats non conseillés
Document texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	CSV, ODS	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Données statistiques	CSV, XML, TXT, RData (suivant les versions)	SAS, RData (suivant les versions)
Images	JPEG, TIFF, PNG	DICOM
Audio	BWF, MXF, Matroska (.mka), FLAC, OPUS	WAVE, MP3, AAC, AIFF, OGG
Video	MXF, MKV	MPEG-4, MPEG-2, AVI, QuickTime (.mov, .qt)
Information géographique	GML, MIF/MID	ESRI Shapefiles, MapInfo, KML
Images géoréférencées	GeoTIFF (.tif, .tiff)	TIFF World File
Raster	ASCII GRID (.asc, .txt)	ESRI GRID

https://facile.cines.fr/ service de validation des formats

Recommended formats

CORNELL UNIVERSITY LIBRARY

File formats for digital content: Probability for full long-term preservation

Content type	High	Medium	Low
Text	• Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM) • XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema) • PDF/A-1 (ISO 19005-1) (*.pdf)	• Cascading Style Sheets (*.css) • DTD (*.dtd) • Plain text (ISO 8859-1 encoding • PDF (*.pdf) (embedded fonts) • Rich Text Format 1.x (*.rtf) • HTML (include a DOCTYPE declaration) • SGML (*.sgml) • Open Office (*.sxw*.odt) • OOXML (ISO/IEC DIS 29500) (*.docx) • Microsoft Word 2007 or newer (*.docx)	• PDF (*.pdf) (encrypted) • Microsoft Word 2003 or older (*.doc) • WordPerfect (*.wpd) • DVI (*.dvi) • All other text formats not listed
Raster image	• TIFF (uncompressed) • JPEG2000 (lossless) (*.jp2) • PNG (*.png)	• BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy) (*.jp2) • TIFF (compressed) • GIF (*.gif) • Digital Negative DNG (*.dng)	• MrSID (*.sid) • TIFF (in Planar format) • FlashPix (*.fpx) • PhotoShop (*.psd) • RAW • JPEG 2000 Part 2 (*.jpf, *.jpx) • All other raster image formats not listed
Vector graphics	• SVG (no Java script binding) (*.svg)	• Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	• Encapsulated Postscript (EPS) • Macromedia Flash (*.swf) • All other vector image formats not listed
Audio	• AIFF (96kHz 16bit PCM) (*.aif, *.aiff) • WAV (96kHz 24bit PCM) (*.wav)	• SUN Audio (uncompressed) (*.au) • Standard MIDI (*.mid, *.midi) • Ogg Vorbis (*.ogg) • Free Lossless Audio Codec (*.flac) • Advance Audio Coding (*.mp4, *.m4a, *.aac) • MP3 (MPEG-1/2, Layer 3) (*.mp3)	• AIFC (compressed) (*.aifc) • NeXT SND (*.snd) • RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) • Windows Media Audio (*.wma) • Protected AAC (*.m4p) • WAV (compressed) (*.wav) • All other audio formats not listed
Video	• Motion JPEG 2000 (ISO/IEC 15444-4)??*.mj2) • AVI (uncompressed/native, motion JPEG) (*.avi) • QuickTime Movie (uncompressed/native, motion JPEG) (*.mov)	• Ogg Theora (*.ogg) • MPEG-1, MPEG-2 (*.mpg, *.mpeg, wrapped in AVI, MOV) • MPEG-4 (H.263, H.264) (*.mp4, wrapped in AVI, MOV)	• AVI (others) (*.avi) • QuickTime Movie (others) (*.mov) • RealNetworks 'Real Video' (*.rv) • Windows Media Video (*.wmv) • All other video formats not listed

Formats : keep in mind …

- Use open formats to facilitate sharing and interoperability
- The format must be documented (standards)

- If you use closed formats, you must ensure sustainability and check whether the conversion alters the information

https://en.wikipedia.org/wiki/Open_file_format

Make your data reusable … the smart way

- As open as possible, as shared as necessary.
- Make it clear

LICENSE CHOOSER

Follow the steps to select the appropriate license for your work. This site does not store any information.

1 **License Expertise**
I need help selecting a license.

2 **Attribution**
Anyone can use my work, even without giving me attribution.

3 **Commercial Use**
Others can use my work, even for commercial purposes.

4 **Derivative Works**
Others can remix, adapt, or build upon my work.

5 **Sharing Requirements**
Others must use the same CC license if they adapt my work.

6 **Confirm that CC licensing is appropriate**
I confirmed the appropriateness of CC licensing.

7 **Attribution Details**

Filling out this form is optional, but helps others attribute your work to you, and fills in machine-readable code.

Title of Work

This work

RECOMMENDED LICENSE

CC BY-SA 4.0

Attribution-ShareAlike 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, even for commercial purposes. If others remix, adapt, or build upon the material, they must license the modified material under identical terms.

BY: Credit must be given to you, the creator.

SA: Adaptations must be shared under the same terms.

See the License Deed

MARK YOUR WORK

Choose the kind of work to get appropriate license code or public domain marking.

Website | **Print Work or Media**

If you are licensing or marking one work, paste the code next to it. If you are licensing or marking the whole page or blog, you can paste the code at the bottom of the page.

Rich Text | HTML | XMP

This work is licensed under CC BY-SA 4.0

license abbreviation | full license name

Copy

LICENSE CHOOSER

Follow the steps to select the appropriate license for your work. This site does not store any information.

1 **License Expertise**
I need help selecting a license.

2 **Attribution**
Anyone can use my work, even without giving me attribution.

3 **Commercial Use**
Others can use my work, even for commercial purposes.

4 **Derivative Works**
Others can remix, adapt, or build upon my work.

5 **Sharing Requirements**
Others must use the same CC license if they adapt my work.

6 **Confirm that CC licensing is appropriate**
I confirmed the appropriateness of CC licensing.

7 **Attribution Details**

Filling out this form is optional, but helps others attribute your work to you, and fills in machine-readable code.

Title of Work

This work

RECOMMENDED LICENSE

CC BY-SA 4.0

Attribution-ShareAlike 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, even for commercial purposes. If others remix, adapt, or build upon the material, they must license the modified material under identical terms.

BY: Credit must be given to you, the creator.

SA: Adaptations must be shared under the same terms.

See the License Deed

MARK YOUR WORK

Choose the kind of work to get appropriate license code or public domain marking.

Website | **Print Work or Media**

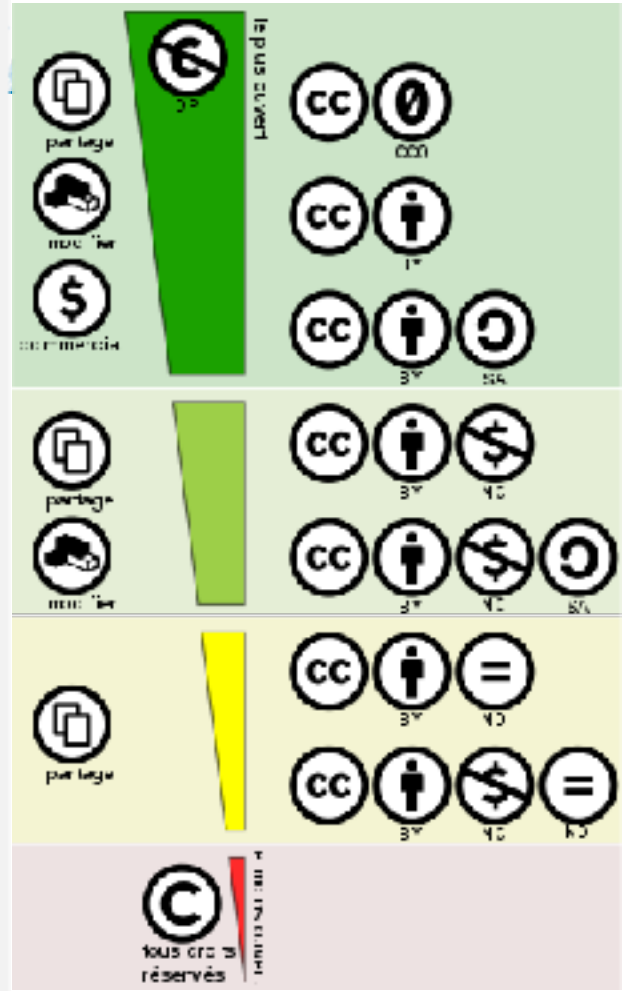
If you are licensing or marking one work, paste the code next to it. If you are licensing or marking the whole page or blog, you can paste the code at the bottom of the page.

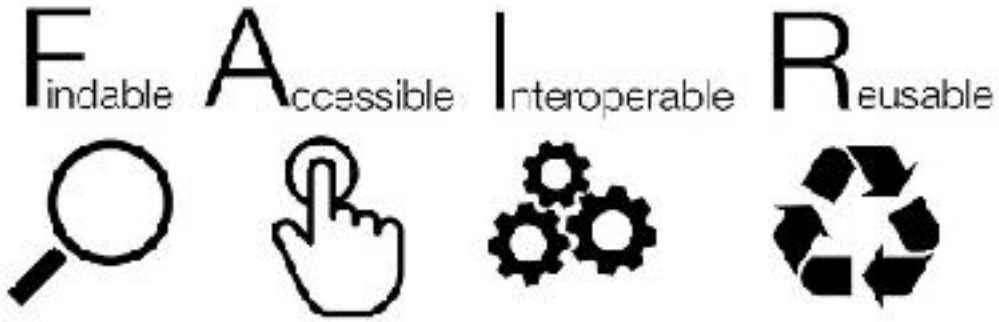
Rich Text | HTML | XMP

This work is licensed under CC BY-SA 4.0

license abbreviation | full license name

Copy





Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

A persistent identifier is assigned to your data
There are rich metadata, describing your data
The metadata are online in a searchable resource e.g. a catalogue or data repository
The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

Following the persistent ID will take you to the data or associated metadata
The protocol by which data can be retrieved follows recognised standards e.g. http
The access procedure includes authentication and authorisation steps, if necessary
Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

Data is provided in commonly understood and preferably open formats
The metadata provided follows relevant standards
Controlled vocabularies, keywords, thesauri or ontologies are used where possible
Qualified references and links are provided to other related data

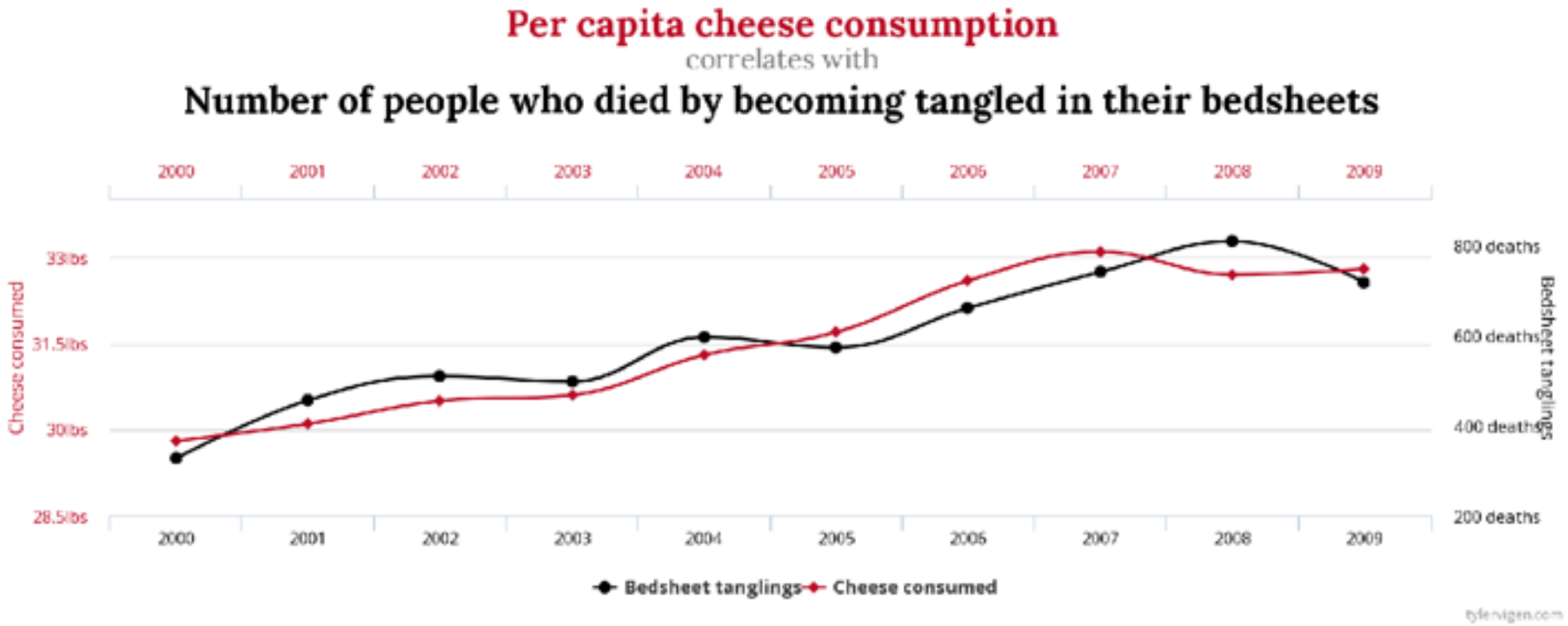
Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

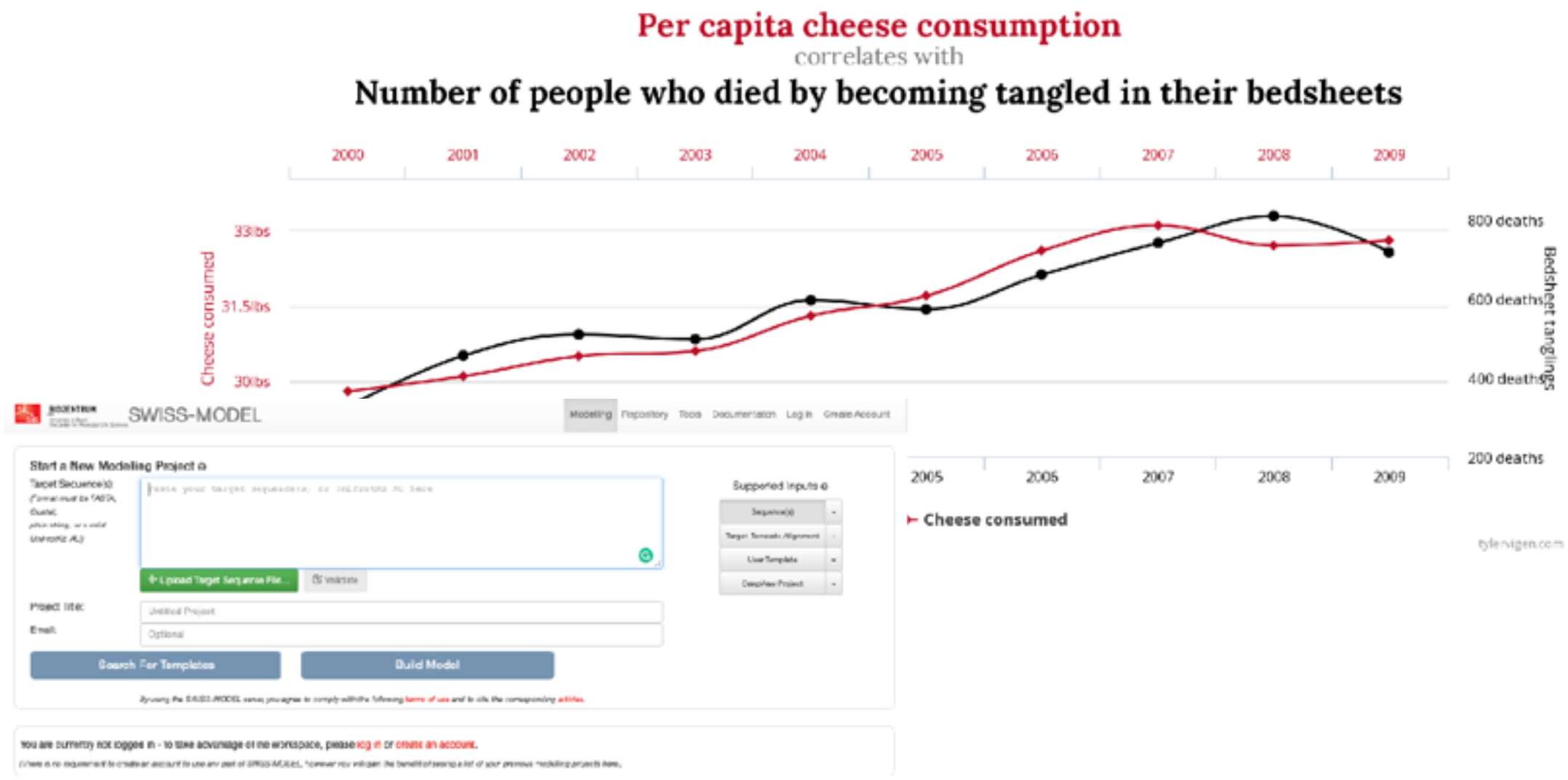
The data are accurate and well described with many relevant attributes
The data have a clear and accessible data usage license
It is clear how, why and by whom the data have been created and processed
The data and metadata meet relevant domain standards

Data quality : stay smart

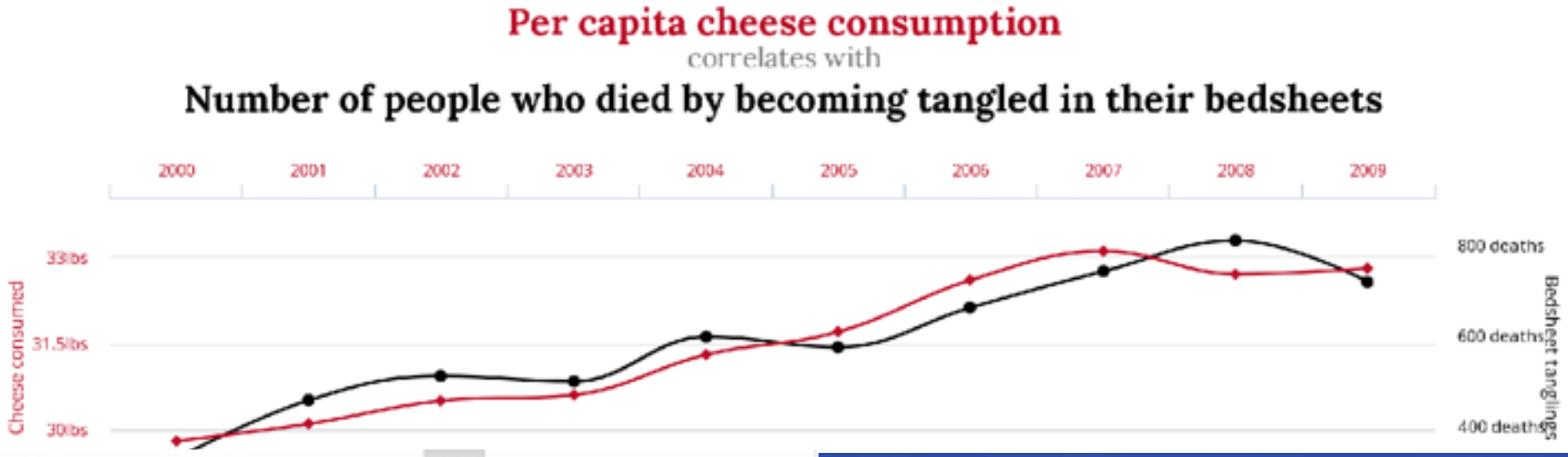
Data quality : stay smart

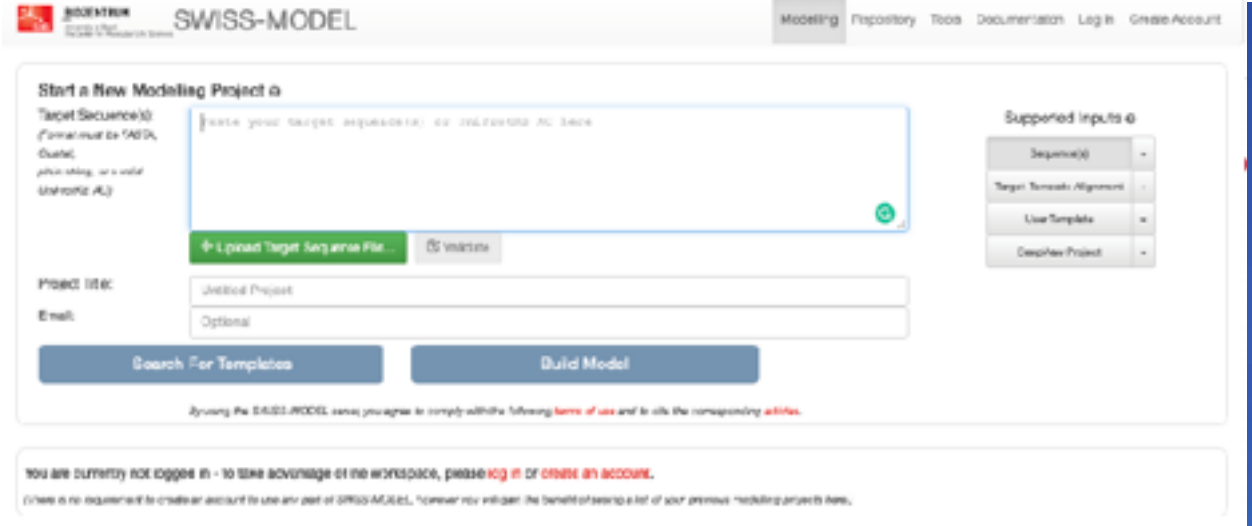


Data quality : stay smart



Data quality : stay smart





Quality check

- Missing data
- Missing values replaced by zero
- Incomplete series
- Duplicated lines or values
- Inconsistent spelling
- Inconsistent date formats (1900, 1904)
- 65 536 lines (or 255 columns)

Quality check

- Missing data
- Missing values replaced by zero
- Incomplete series
- Duplicated lines or values
- Inconsistent spelling
- Inconsistent date formats (1900, 1904)
- 65 536 lines (or 255 columns)

Open Refine

choice	count
Andy Anderson	
Andy R. Anderson	
Anderson, Andy	
Beatrice Beaufort	
Beatrice Mansfield	
Beaufort, Beatrice	
...	...

Resolving some problems

- Encoding
- CR / LF
- « PDFied » Data
- Scanned Data

Resolving some problems

- Encoding
- CR / LF
- « PDFied » Data
- Scanned Data

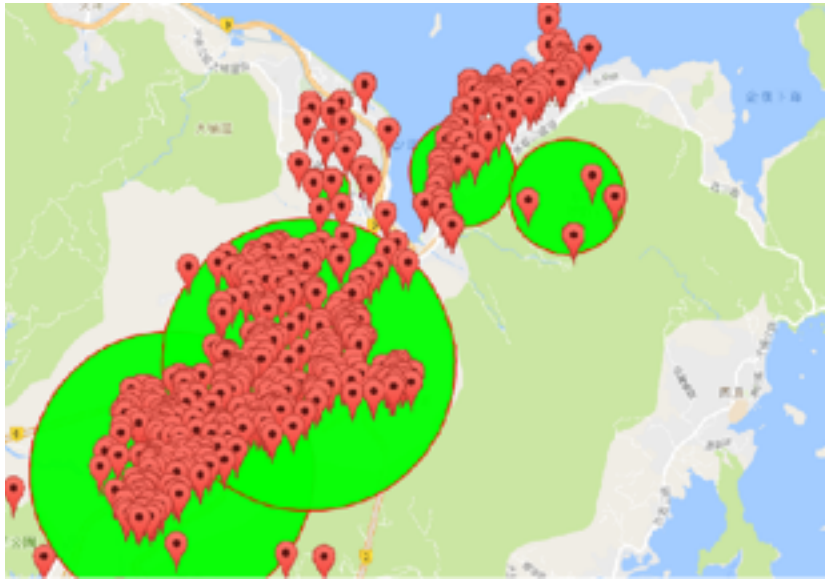
Tabula

Or just give up !

- Unreliable source
- Transparency of the collection process
- Unrealistic data accuracy
- Unexplainable outliers

Or just give up !

- Unreliable source
- Transparency of the collection process
- Unrealistic data accuracy
- Unexplainable outliers



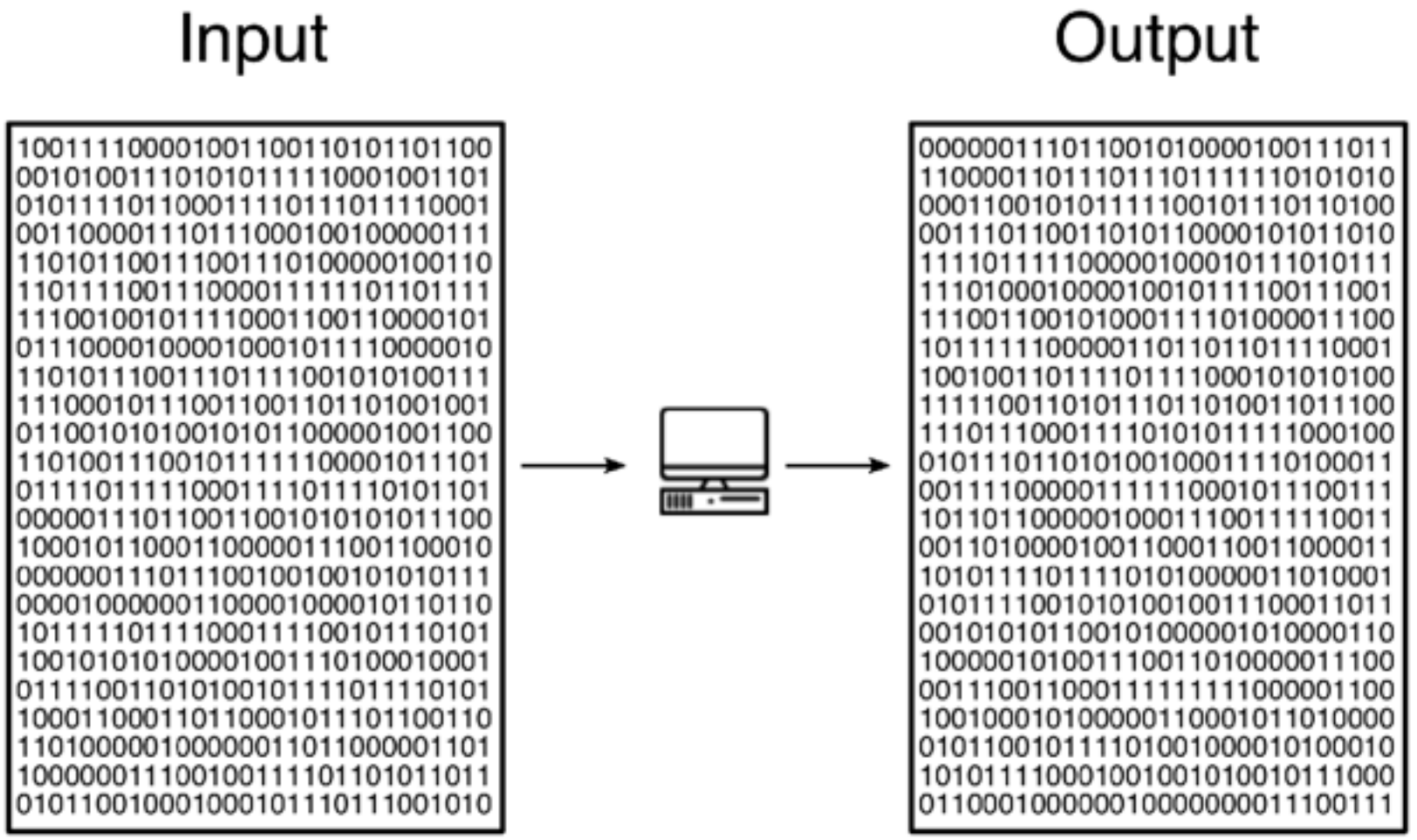
Data Analysis

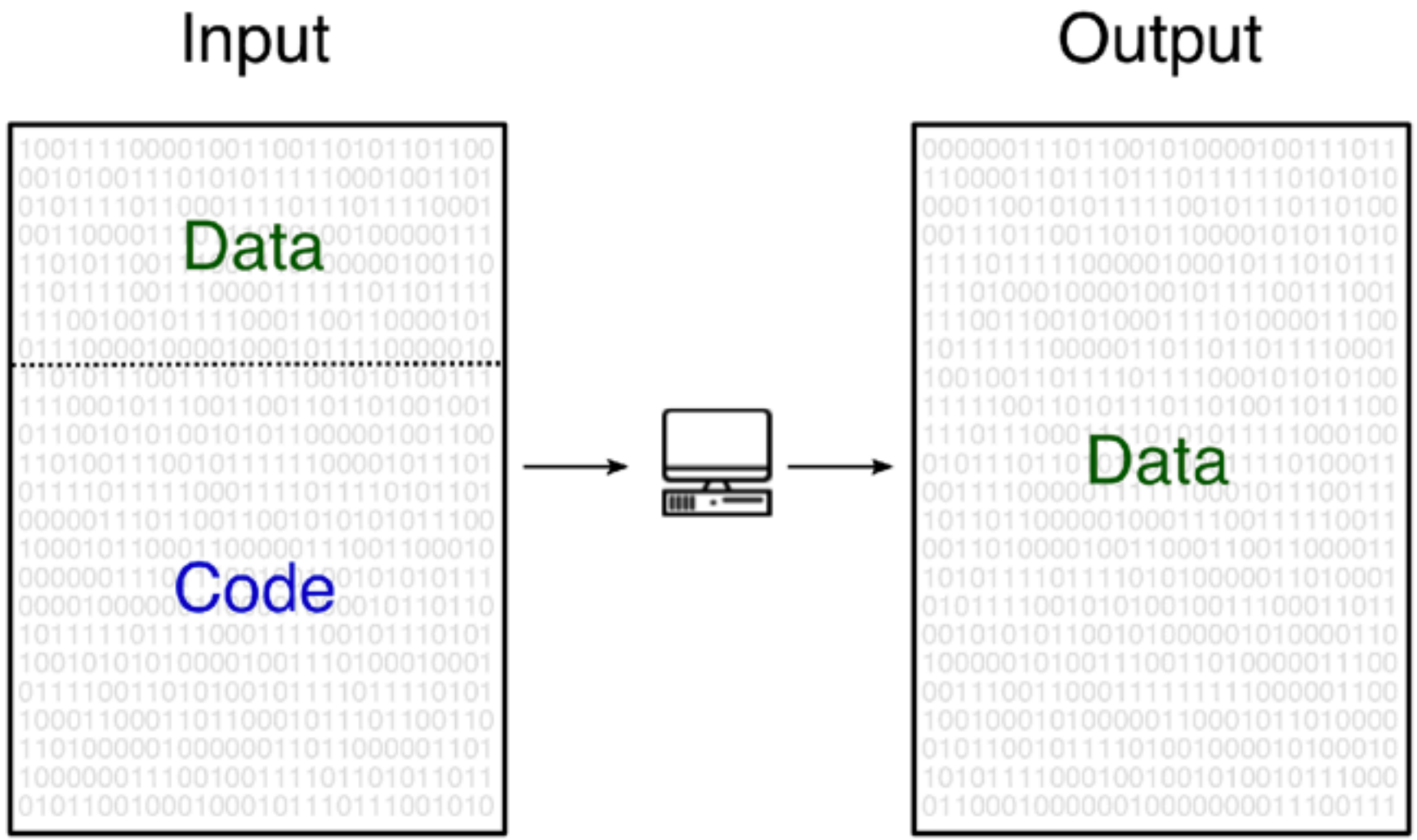
Every Scientist is a DATA scientist

J’peux le refaire ?

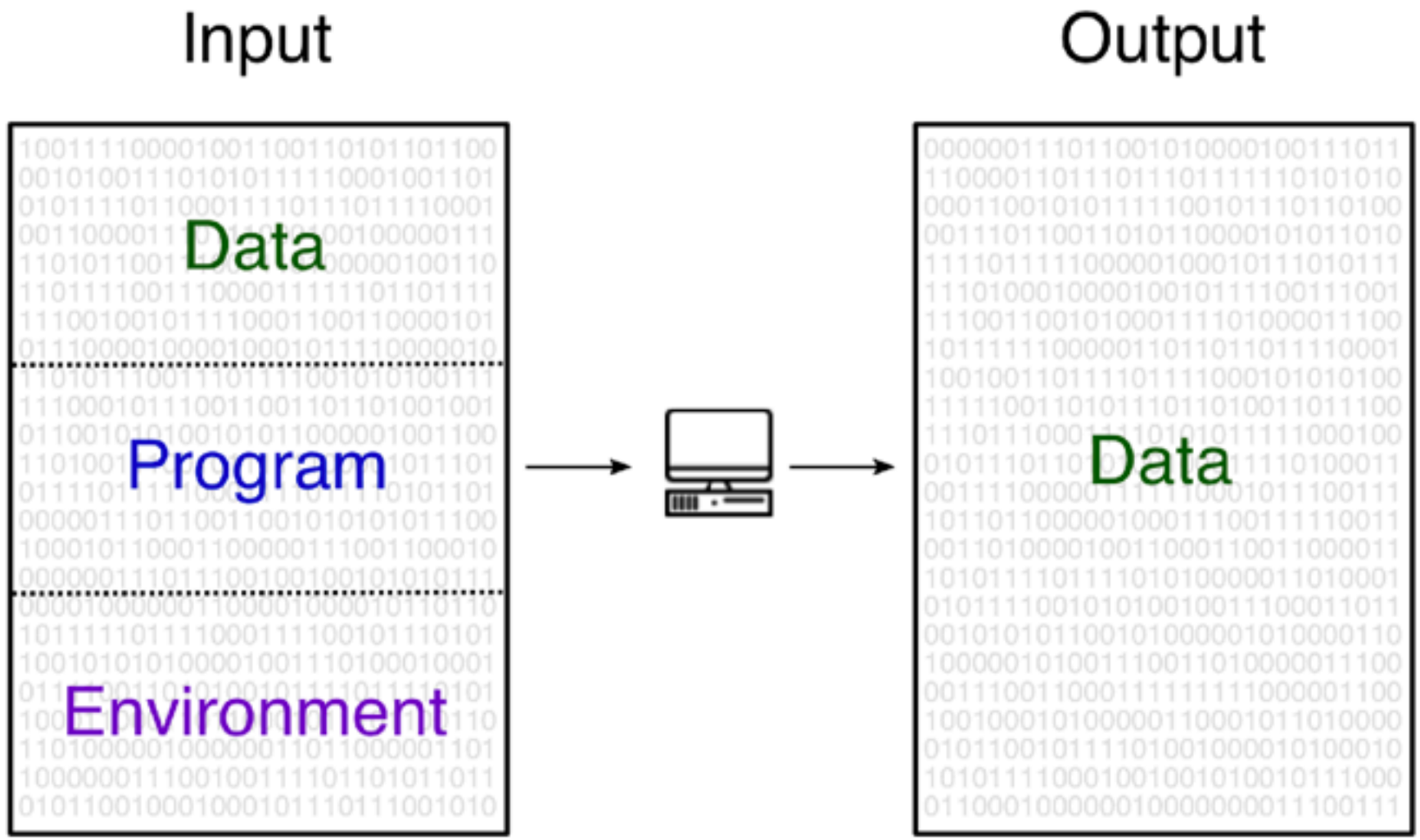
J’peux le refaire ?

Data Analysis





Computer by Creative Stall from the Noun Project



Computer by Creative Stall from the Noun Project

Not that !

my research

my colleagues' code

stuff I don't care about

ONCOSPHERE
International Meeting

12-13 janvier 2023

Bordeaux

ONCOSPHERE

What remains too be done ?

BMC Medecine (Hamilton *et al.* 2022)

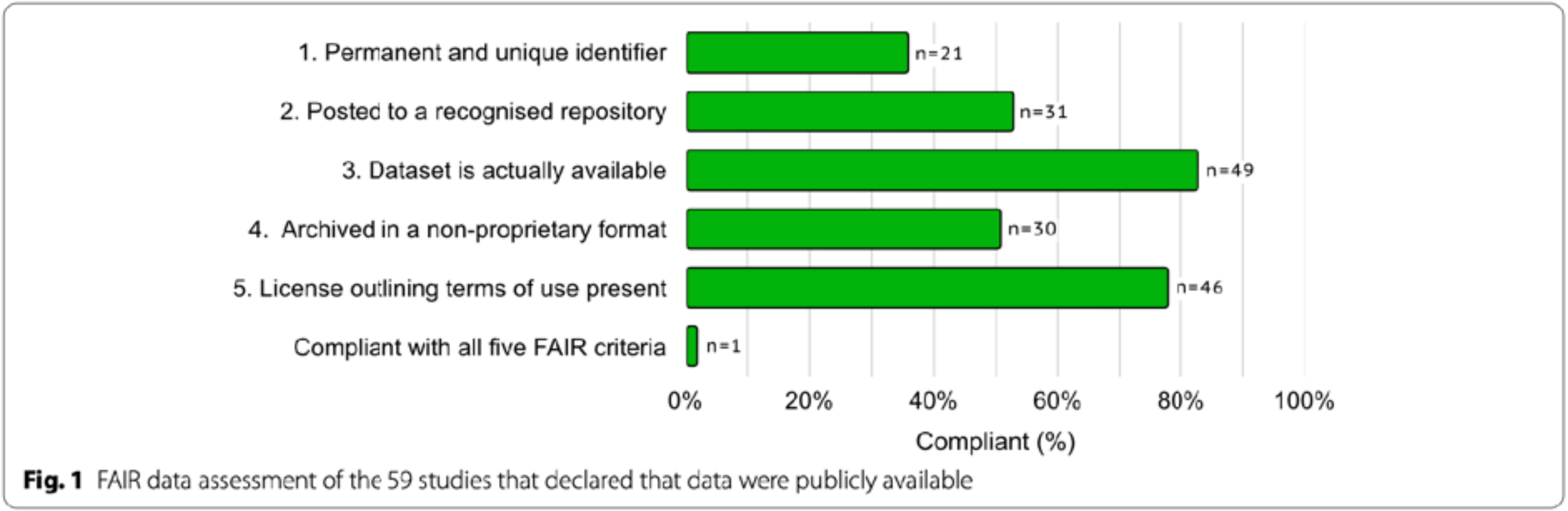
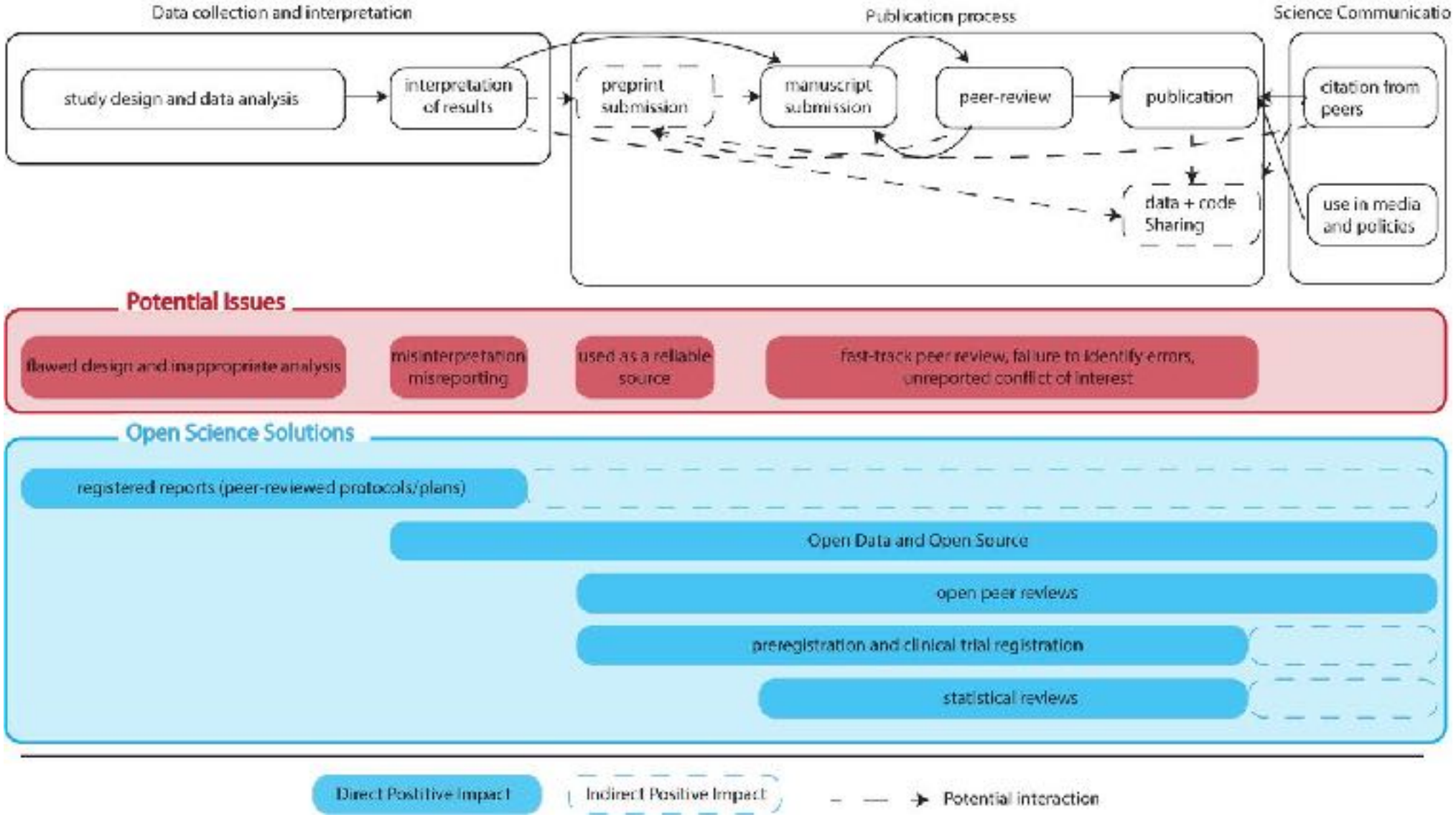


Fig. 1 FAIR data assessment of the 59 studies that declared that data were publicly available

Open Science Saves Lives



A success story

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

The key to the Alzheimer's project was an **agreement to share all the data**, making **every single finding public immediately**, available to anyone with a computer anywhere in the world. **No one would own the data**. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

The key to the Alzheimer's project was an **agreement to share all the data**, making **every single finding public immediately**, available to anyone with a computer anywhere in the world. **No one would own the data**. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

"It's not science the way most of us have practiced it in our careers. But we all realized that we would never get biomarkers unless all of us parked our egos and intellectual-property noses outside the door and agreed that all of our data would be public immediately."

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

The key to the Alzheimer's project was an **agreement to share all the data**, making **every single finding public immediately**, available to anyone with a computer anywhere in the world. **No one would own the data**. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

"It's not science the way most of us have practiced it in our careers. But we all realized that we would never get biomarkers unless all of us parked our egos and intellectual-property noses outside the door and agreed that all of our data would be public immediately."

At first, the collaboration struck many scientists as worrisome — they would be giving up ownership of data, and anyone could use it, publish papers, maybe even misinterpret it and publish information that was wrong.

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

The key to the Alzheimer's project was an **agreement to share all the data**, making **every single finding public immediately**, available to anyone with a computer anywhere in the world. **No one would own the data**. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

"It's not science the way most of us have practiced it in our careers. But we all realized that we would never get biomarkers unless all of us parked our egos and intellectual-property noses outside the door and agreed that all of our data would be public immediately."

At first, the collaboration struck many scientists as worrisome — they would be giving up ownership of data, and anyone could use it, publish papers, maybe even misinterpret it and publish information that was wrong.

Companies as well as academic researchers are using the data. There have been more than **3,200 downloads** of the entire massive data set and almost a **million downloads of the data sets containing images from brain scans**.

A success story

Sharing of Data Leads to Progress on Alzheimer's NYT By GINA KOLATA Published: August 12, 2010

In 2003, a group of scientists and executives from the **NIH**, the **FDA**, the **drug and medical-imaging industries**, **universities** and **nonprofit groups** joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

The key to the Alzheimer's project was an **agreement to share all the data**, making **every single finding public immediately**, available to anyone with a computer anywhere in the world. **No one would own the data**. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

"It's not science the way most of us have practiced it in our careers. But we all realized that we would never get biomarkers unless all of us parked our egos and intellectual-property noses outside the door and agreed that all of our data would be public immediately."

At first, the collaboration struck many scientists as worrisome — they would be giving up ownership of data, and anyone could use it, publish papers, maybe even misinterpret it and publish information that was wrong.

Companies as well as academic researchers are using the data. There have been more than **3,200 downloads** of the entire massive data set and almost a **million downloads of the data sets containing images from brain scans**.

"We weren't sure, frankly, how it would work out having data available to everyone," he said. "But we felt that the good that could come out of it was overwhelming. And that's what's happened."

The place to go : RDMkit

RDMkit

Home About Contribute Contact GitHub Search RDMkit

Data life cycle

Your role

Your comain

Your tasks

Tool assembly

All tools and resources

All training resources

Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing good data management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, the RDMkit has guidelines, information, and pointers to help you with problems throughout the data's life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable and Reusable — by-design, from the first steps of data management planning to the final steps of depositing data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

Data life cycle

Start here to get an overview of research data management. Click on a section of the diagram below to get an introduction to that stage of the data management life cycle.

Data life cycle

Your role

Your domain

Your tasks

Tool assembly

All tools and resources

All training resources

Preserving

Edit this page

- What is data preserving?
- Why is data preserving important?
- What should be considered for preserving data?
- More information

What is data preserving?

Data preservation consists of a series of activities necessary to ensure safety, integrity and accessibility of data for as long as necessary, even decades. Data preservation is indeed more than just data storage and backup, since data can be stored and backed up without being preserved. Data preservation prevents data from becoming unavailable and unusable over time through appropriate activities, such as:

- Ensure data safety and integrity.
- Change the file format (format migration) and update software to make sure that they do not become outdated or obsolete.
- Change hardware and other storage media (such as paper, magnetic tape, etc) to avoid degradation.

Data life cycle

Your role

Your domain

Your tasks

Tool assembly

All tools and resources

All training resources

Sharing

Edit me

- What is data sharing?
- Why is data sharing important?
- What should be considered for data sharing?
- More information

What is data sharing?

Sharing data means making your data known to other people.

You can share your data with collaboration partners in the context of a collaborative research project, or you can publish your data to share it with the global research community and society at large.

It's important to know that data sharing doesn't mean open data or public data. You can choose to share your data with restricted access or even closed access. Moreover, sharing or publishing data is different from publishing a paper or a manuscript in a journal. Here we focused on data (i.e. raw observations and measurements, analysis workflows, code, etc), not on papers or articles.

Data sharing can be done at any time during the research data life cycle but, at the latest, data should be made available at the time of publication of articles that use the data to make scientific conclusions.

A remporter à la maison

DMP

https://easydmp.eudat.eu/plan/

https://pgd17juin2019.sciencesconf.org/program

Data Repositories

https://fairsharing.org/databases/

https://www.nature.com/sdata/policies/repositories

http://about.zenodo.org

https://search.datacite.org/data-centers

https://cordis.europa.eu/project/rcn/207500/

Data Management

https://www.csc.fi/en/web/training/-/webinar-research-data-management

http://www.smalsresearch.be/dix-bonnes-pratiques-pour-ameliorer-et-maintenir-la-qualite-des-donnees/

https://mantra.edina.ac.uk

https://www.fairdata.fi/en/

Licences

https://www6.inra.fr/datapartage/Partager-Publier/Choisir-une-licence

https://creativecommons.org/choose/

Ontologies &Metadata Standards

https://fairsharing.org/databases/

https://bioportal.bioontology.org

https://isa-tools.org

https://datacite.org

And also

https://eudat.eu

http://www.dcc.ac.uk

https://github.com/laurence001/quartz-bad-data-guide-french

https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html

https://doranum.fr/enjeux-benefices/parcours-interactif-sur-la-gestion-des-donnees-de-la-recherche_10_13143_3xnz-as06/

INRAE

WISHES YOU
A HAPPY NEW YEAR
2023

All the best for 2023 & Thank you for your attention