

Building a toolkit for FAIR data and Open science

for the Computational Biology Training in Hematology programme, EHA

Kim Ferguson, DANS-KNAW

13/06/2024 – EHA2024 Congress

Brief Introduction

What is DANS?

- DANS is an institute of the KNAW (Royal Netherlands Academy of Arts and Sciences) and the NWO (Dutch Research Council)
- DANS is the Dutch national centre of expertise and repository for research data
- DANS maintains four domain-specific repositories (Data Stations), supports institutional repositories (DataverseNL), and works on various national and international data projects

Who am I?

- Kim Ferguson (PhD in Insect Genomics)
- Research Data Management Specialist → European projects, training, FAIR data

The background is a solid blue color. There are several white geometric shapes: a large arc at the top left, a small circle on the left side, and a large arc at the bottom center.

Refresher on FAIR data

⋮

Flashback to May...

Open science and making data FAIR

Open Science in Life Science/Medical Research

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 Segalas⁵, Haiting Jiang⁶, Paola Masuzzo⁷, Cooper Smout⁷, Eric Billy⁸, Maxime Deforet⁹ and Clémence Leyrat^{5,10}



Clinical and Translational Oncology
<https://doi.org/10.1007/s12094-024-03468-7>

RESEARCH ARTICLE

State of open science in cancer research

Cristina Rius^{1,2,3,4} · Yiming Liu^{1,2} · Andrea Sixto-Costoya^{1,2,5} · Juan Carlos Valderrama-Zurián^{1,2} · Rut Lucas-Dominguez^{1,2,6} 

Received: 2 February 2024 / Accepted: 15 March 2024

JAMIA Open, 3(3), 2020, 472–486
doi: 10.1093/jamiaopen/oaaa030
Advance Access Publication Date: 11 September 2020
Review



Review

The case for open science: rare diseases

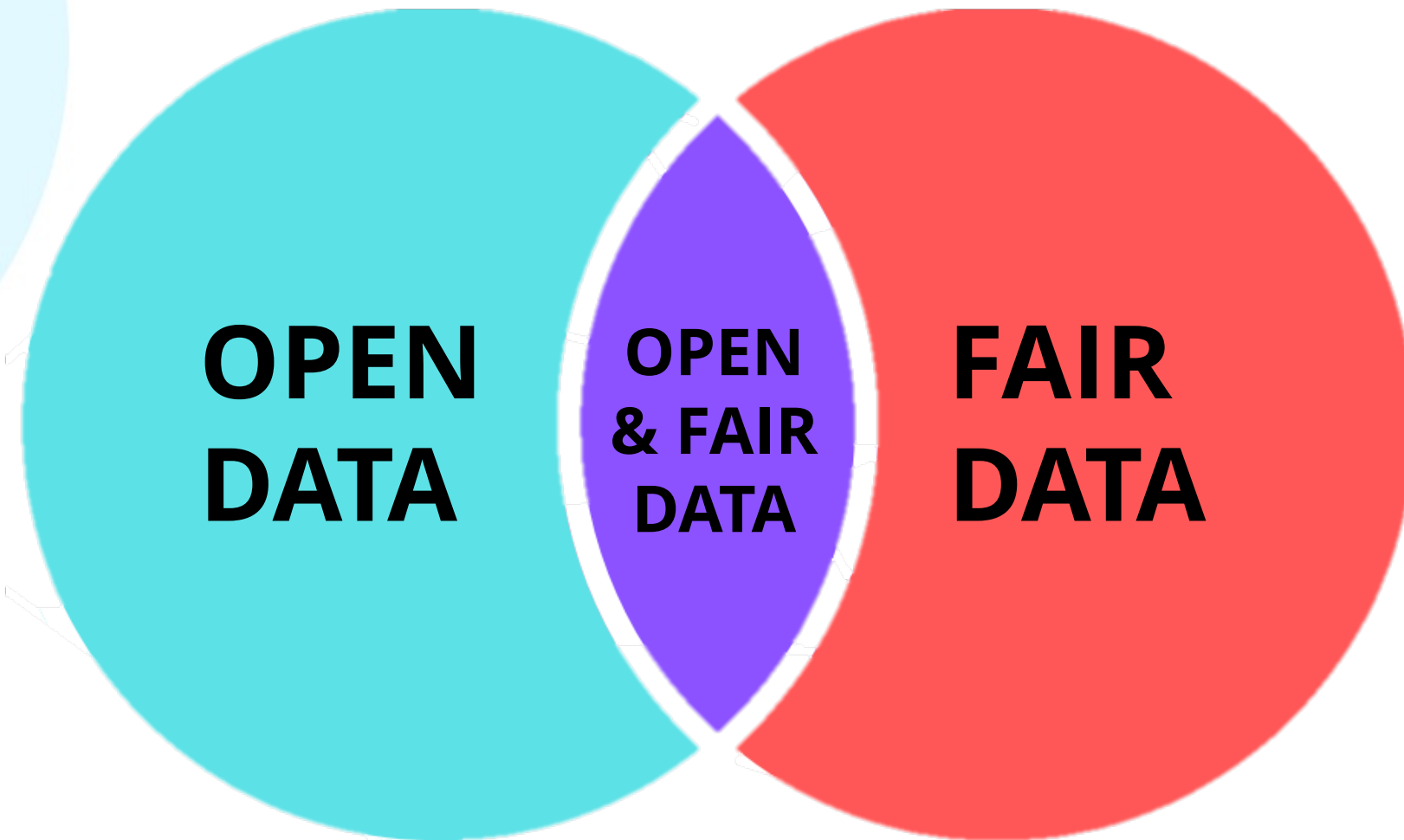
Yaffa R. Rubinstein,¹ Peter N. Robinson,² William A. Gahl,³ Paul Avillach,⁴ Gareth Baynam,⁵ Helene Cederroth,⁶ Rebecca M. Goodwin,⁷ Stephen C. Groft,⁸ Mats G. Hansson,⁹ Nomi L. Harris,¹⁰ Vojtech Huser,¹¹ Deborah Mascalzoni,¹² Julie A. McMurry,¹³ Matthew Might,¹⁴ Christoffer Nellaker,¹⁵ Barend Mons,¹⁶ Dina N. Paltoo,⁷ Jonathan Pevsner,¹⁷ Manuel Posada,¹⁸ Alison P. Rockett-Frase,¹⁹ Marco Roos,²⁰ Tamar B. Rubinstein,²¹ Domenica Taruscio,²² Esther van Enkevort¹⁹,²³ and Melissa A. Haendel¹³

Findable, Accessible, Interoperable, and Reusable

- First publication, “The FAIR Guiding Principles for scientific data management and stewardship” from Wilkinson et al. 2016
- Continues to be enhanced
 - FAIR principles and hardware (Miljković, Trisovic, and Peer 2021)
 - FAIR principles and software (Barker et al. 2022)
 - FAIR principles and health research management outcomes (Martínez-García et al. 2023)
- Referenced in national and international policies and funding agreements



There is some overlap





“As open as possible, as closed as necessary”
“If it can’t be open, it should at least be FAIR”

Health data, sensitive data – sometimes, this cannot be open.

However, metadata can nearly always be open

- What was measured, how it was measured, types of statistical analyses, etc.
- Tricky area? IPR → but compromises are possible



Either way, your data can still be FAIR



Relying on search engines for findability...

Google Search Really Has Gotten Worse, Researchers Find



JASON KOEBLER · JAN 16, 2024 AT 1:28 PM

~~404~~


- Study used is [Bevendorff and Wiegmann et al. 2024](#) – this [404 Media article](#) contains a link to a PDF of the article, a longitudinal study over a year (Aug. 2022 – Sept. 2023)
- Focused on product reviews, but the listed approaches to SEO (search engine optimisation) apply to most web-based content and website structuring
- “The constant struggle of billion-dollar search engine companies with targeted SEO affiliate spam should serve as an example that web search is a dynamic game with many players, some with bad intentions.”

Tools for Open & FAIR data

•
•
•
•
•
•

#1 ORCID





<https://orcid.org/0000-0001-6483-1936>

Emails >
kim.ferguson@dans.knaw.nl

Websites & social links >
ResearchGate
[Royal Netherlands Academy of Arts & Sciences \(KNAW\) profile page](#)

Other IDs >
Digital Author ID: 426283309
Scopus Author ID: 57194440509

Countries >
Netherlands

 Printable version

Published name
Kim B Ferguson

Name
Kim Ferguson

Activities [Expand all](#)



> **Employment (5)** [Sort](#)

> **Education and qualifications (1)** [Sort](#)

▼ **Works (22)** [Sort](#)


Bracon brevicornis Genome Showcases the Potential of Linked-Read Sequencing in Identifying a Putative Complementary Sex Determiner Gene

Genes
2020-11-24 | Journal article
DOI: [10.3390/genes11121390](#)
CONTRIBUTORS: Kim B. Ferguson; Bart A. Pannebakker; Alejandra Centurión; Joost van den Heuvel; Ronald Nieuwenhuis; Frank F. M. Becker; Elio Schijlen; Andra Thiel; Bas J. Zwaan; Eveline C. Verhulst
[Show more detail](#)

Source:  Crossref  Preferred source (of 2)


DANS Data Trail "Engaging with the CESSDA Data Archiving Guide"

2023-05-16 | Other | Author
SOURCE-WORK-ID: 49af5eb0-9c18-435a-b7d9-677c65abe96b
CONTRIBUTORS: Maaik Verburg; Kim B Ferguson; Ricarda Braukmann; Valentijn Gilissen; Simon Saldner
[Show more detail](#)

Source:  Royal Netherlands Academy of Arts and Science (KNAW)

DARIAH Annual Event 2022 - Storytelling

DARIAH Annual Event 2022 Storytelling, Athens, Greece, 31/05/2022
2022 | Edited book | Editor
DOI: [10.5281/zenodo.6720075](#)
SOURCE-WORK-ID: d3d5d335-cdb7-4066-a379-4266b8f088b7
CONTRIBUTORS: Jennifer Edmond; Andrea Scharnhorst; Francesca Morselli; Agiatis Benardou; Eliza Papaki; Kim B Ferguson
[Show more detail](#)

Source:  Royal Netherlands Academy of Arts and Science (KNAW)

- A persistent identifier for researchers
- Tracks you beyond a single position
- Aometimes a requirement for publication in certain journals
- Can help track publications, datasets, and other materials archived or published

Do you have an **ORCID ID** yet?

Data repositories: good for finding data,
good for storing data

Internal shared drives are useful for
colleagues but aren't the same as publishing
a dataset in a repository.

“A flash drive is not a storage solution”

Repositories & FAIR

A good repository:

- Provides a Persistent Identifier (PID) to your dataset = **Findable**
- Enables the inclusion of rich metadata = **Reusable**
- Allows for human and machine findable and readable metadata = **Findable** + **Accessible**
- Uses Knowledge Organisation Systems (KOS), such as vocabularies, to improve the quality of metadata = **Reusable**
- Provides long term and easy access to your data = **Accessible** + **Reusable**

#2 Re3data

Re3data can help with a few questions:

- Where to find data to re-use?
- Where to store data of a certain type?
- What is the certification level of some of these repositories
- What standards should I be aware of in my field?



What questions re3data cannot answer:

- Is this the most appropriate place to store or find data?
- Is this “good” data
- Why isn’t my favourite repository in re3data?

What if it's not in re3data?

Reasons why it may not be in re3data:

- It hasn't been added yet, it's a hidden gem!
- It isn't currently meeting the standards of re3data
- It hasn't been added yet, it's too new!



[Database \(Oxford\)](#). 2022; 2022: baac003.

Published online 2022 Mar 9. doi: [10.1093/database/baac003](https://doi.org/10.1093/database/baac003)

PMCID: PMC9216516

PMID: [35262674](https://pubmed.ncbi.nlm.nih.gov/35262674/)

ImmuneData: an integrated data discovery system for immunology data repositories

[Nan Deng](#), [Canglin Wu](#), [Ashraf Yaseen](#), and [Hulin Wu](#)[✉]

Sidebar: Repository certification

Certification relates to a few elements:

- Organisational infrastructure (funding, staffing, policies)
- Handling of digital objects (persistent identifiers, documentation)
- Technological infrastructure (IT, security)

In some communities, there is self-regulation or tradition rather than certification. A lack of certification doesn't necessarily mean it's a poor repository. Always check!

Case in point:

“The International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.” (Source: [GenBank](#))

GenBank and the ENA are good repositories:

- Provides an **Accession ID and/or Project ID** as a Persistent Identifier (PID) for your dataset
- Enables the inclusion of rich metadata, **including minimum required information for submission**
- Allows for human and machine findable and readable metadata
- Uses **several controlled vocabularies** to improve the quality of metadata
- Provides long term and easy access to your data, **shared between multiple organisations**

#3 FAIR-Aware



Your first step towards your FAIR data(set)

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

- Just 10 questions
- Self-paced, supporting documentation
- Related to the FAIR-ness of a dataset
- Can be used just before publication or shortly after for improving FAIR qualities
- Can also be done to determine how to improve previous publications

#4 & #5 Figshare & Zenodo

Both useful for depositing and sharing things beyond publications

- Supplementary information for publications
- Conference presentations
- Posters
- Infographics
- Videos
- Project deliverables (reports, data management plans)
- Data is possible, but not recommended - use a domain-specific repository for that (trust me)

#6 Elixir RDM Cookbook

A well-curated home for resources and tips for research data management, including FAIR practices

- Role-specific resources: [Researcher](#)
- Resources for each stage of the [research data life cycle](#)
- Fine-tune advice to your [research domain](#)
- [Country-specific](#) pointers and contacts

But where to start with Open Science and FAIR data? A few tips

- [University of Oxford](#)
- [University of Cambridge](#)
- [University of Glasgow](#)
- [University of Manchester](#)
- [Amsterdam Medical Centre](#) (incl. [FAIR Data](#))
- [UMC Groningen](#)
- [Open Science Community Rotterdam](#)
- [UGent Open Science](#)

- [Université Paris Cité](#)
- [Université de Lille](#)
- [University of Western Australia](#)
- [Stanford Center for Open and REproducible Science](#)
- UKSH & [Lübeck Open Science Initiative](#)
- [Frankfurt Open Science Initiative](#)
- [Open Science Center Cologne](#)
- [Repositorio Institucional de la Universidad de Oviedo](#)

*Note: This is the best option I found, but it could be that there are better, more specific-/department-based options

But where to start with Open Science and FAIR data? A few more tips

More for the domain/field that you're working in ERICs (European Research Infrastructure Consortium)

- ELIXIR Europe ERIC, [nodes](#)
 - [RDMKit](#) (Research Data Management Kit)
- The EATRIS ERIC (translational medicine)
- BBMRI-ERIC (Biobanking and BioMedical Research resources Infrastructure) and their [ELSI \(Ethics, Legal, Social Issues\) Services](#)

Other EU Resources

- GDPR.EU including [What is personal data?](#)
- EU Ethics & Data Protection [Decision Tree](#)

The background is a solid blue color. There are several white geometric shapes: a large arc at the top left, a full circle on the left side, and a large arc at the bottom center.

Questions?

⋮

More information

Visit our website www.dans.knaw.nl

And follow us online



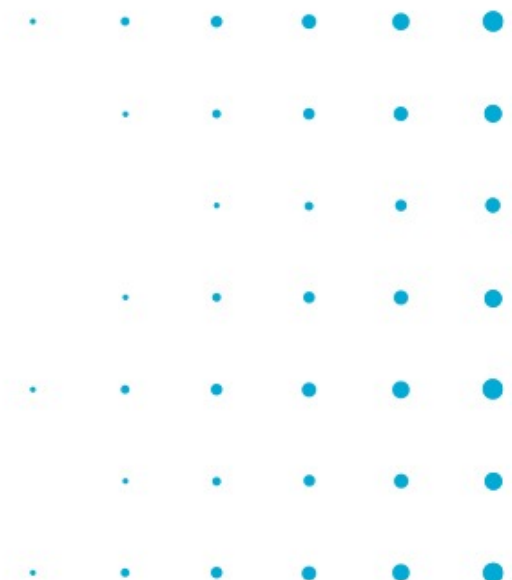
Mastodon
[@DANS_knaw_nwo](#)



LinkedIn
[@DANS](#)



X
[@DANS_knaw_nwo](#)



References (articles & materials)

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, et al. 2022. 'Introducing the FAIR Principles for Research Software'. *Scientific Data* 9 (1): 622. <https://doi.org/10.1038/s41597-022-01710-x>.

Besaçon, Lonni, Nathan Peiffer-Smadja, Corentin Segalas, Haiting Jiang, Paola Masuzzo, Cooper Smout, Eric Billy, Maxime Deforet, and Clémence Leyrat. 2021. 'Open Science Saves Lives: Lessons from the COVID-19 Pandemic'. *BMC Medical Research Methodology* 21 (1): 117. <https://doi.org/10.1186/s12874-021-01304-y>.

Bevendorff, Janek, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. 'Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines'. In *Advances in Information Retrieval*, edited by Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, 56–71. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56063-7_4.

Deng, Nan, Canglin Wu, Ashraf Yaseen, and Hulin Wu. 2022. 'ImmuneData: An Integrated Data Discovery System for Immunology Data Repositories'. *Database: The Journal of Biological Databases and Curation* 2022 (March):baac003. <https://doi.org/10.1093/database/baac003>.

Groeneweg, Stefan, Robin P. Peeters, Carla Moran, Athanasia Stoupa, Françoise Auriol, Davide Tonduti, Alice Dica, et al. 2019. 'Effectiveness and Safety of the Tri-Iodothyronine Analogue Triac in Children and Adults with MCT8 Deficiency: An International, Single-Arm, Open-Label, Phase 2 Trial'. *The Lancet Diabetes & Endocrinology* 7 (9): 695–706. [https://doi.org/10.1016/S2213-8587\(19\)30155-X](https://doi.org/10.1016/S2213-8587(19)30155-X).

Heijden, J. a. G. van der, A. J. Kalkdijk-Dijkstra, J. P. E. N. Pierie, H. L. van Westreenen, P. M. A. Broens, B. R. Klarenbeek, and On behalf of the FORCE trial Group. 2022. 'Pelvic Floor Rehabilitation After Rectal Cancer Surgery: A Multicenter Randomized Clinical Trial (FORCE Trial)'. *Annals of Surgery* 276 (1): 38. <https://doi.org/10.1097/SLA.0000000000005353>.

Heijden, J.A.G. Van Der, B.R. Klarenbeek, M. De Vries, P.M. Broens, H.L. Van Westreenen, A.J. Kalkdijk-Dijkstra, and J.P.E.N. Pierie. 2022. 'Pelvic Floor rehabilitation to improve functional Outcome and quality of life after surgery for Rectal Cancer: a randomized controlled trial. FORCE-trial'. Application/pdf,sps,sav. Data Archiving and Networked Services (DANS). <https://doi.org/10.17026/DANS-ZHE-RRV2>.

Hrynaskiewicz, Iain, Melissa L. Norton, Andrew J. Vickers, and Douglas G. Altman. 2010. 'Preparing Raw Clinical Data for Publication: Guidance for Journal Editors, Authors, and Peer Reviewers'. *BMJ* 340 (January):c181. <https://doi.org/10.1136/bmj.c181>.

G. Altman. 2010. 'Preparing Raw Clinical Data for Publication: Guidance for Journal Editors, Authors, and Peer Reviewers'. *BMJ* 340 (January):c181. <https://doi.org/10.1136/bmj.c181>.

References (articles & materials), cont.

Jaarsveld, C.H.M. Van, D.F.M. Reukers, R.P. Akkermans, S.P. Keijmel, G. Morroy, A.S.G. Van Dam, P.C. Wever, et al. 2021. 'Impact of Q-Fever on Physical and Psychosocial Functioning until 8 Years after Coxiella Burnetii Infection'. Application/pdf, csv. Data Archiving and Networked Services (DANS). <https://doi.org/10.17026/DANS-ZPA-FKPH>.

Kalkdijk-Dijkstra, A.J., J.A.G. van der Heijden, H.L. van Westreenen, P.M.A. Broens, M. Trzpis, J.P.E.N. Pierie, B.R. Klarenbeek, et al. 2020. 'Pelvic Floor Rehabilitation to Improve Functional Outcome and Quality of Life after Surgery for Rectal Cancer: Study Protocol for a Randomized Controlled Trial (FORCE Trial)'. *Trials* 21 (1): 112. <https://doi.org/10.1186/s13063-019-4043-7>.

Kraaikamp, Emilie. 2021. 'GDPR for Researchers – Making Your Data Management GDPR Proof [Workshop]', June. <https://doi.org/10.5281/zenodo.5018482>.

———. 2023. 'Health Data for Research - GDPR Challenges', November. <https://doi.org/10.5281/zenodo.10077813>.

Lawlor, Rita T. 2023. 'The Impact of GDPR on Data Sharing for European Cancer Research'. *The Lancet Oncology* 24 (1): 6–8. [https://doi.org/10.1016/S1470-2045\(22\)00653-2](https://doi.org/10.1016/S1470-2045(22)00653-2).

Martínez-García, Alicia, Celia Alvarez-Romero, Esther Román-Villarán, Máximo Bernabeu-Wittel, and Carlos Luis Parra-Calderón. 2023. 'FAIR Principles to Improve the Impact on Health Research Management Outcomes'. *Heliyon* 9 (5): e15733. <https://doi.org/10.1016/j.heliyon.2023.e15733>.

Miljković, Nadica, Ana Trisovic, and Limor Peer. 2021. 'Towards FAIR Principles for Open Hardware', September. <https://doi.org/10.5281/zenodo.5524415>.

Reukers, Daphne F. M., Cornelia H. M. van Jaarsveld, Reinier P. Akkermans, Stephan P. Keijmel, Gabriella Morroy, Adriana S. G. van Dam, Peter C. Wever, et al. 2022. 'Impact of Q-Fever on Physical and Psychosocial Functioning until 8 Years after Coxiella Burnetii Infection: An Integrative Data Analysis'. *PLOS ONE* 17 (2): e0263239. <https://doi.org/10.1371/journal.pone.0263239>.

Rius, Cristina, Yiming Liu, Andrea Sixto-Costoya, Juan Carlos Valderrama-Zurián, and Rut Lucas-Dominguez. 2024. 'State of Open Science in Cancer Research'. *Clinical and Translational Oncology*, April. <https://doi.org/10.1007/s12094-024-03468-7>.

Rubinstein, Yaffa R, Peter N Robinson, William A Gahl, Paul Avillach, Gareth Baynam, Helene Cederroth, Rebecca M Goodwin, et al. 2020. 'The Case for Open Science: Rare Diseases'. *JAMIA Open* 3 (3): 472–86. <https://doi.org/10.1093/jamiaopen/ooaa030>.

Stubbs, Matthew J., Paul Coppo, Chris Cheshire, Agnès Veyradier, Stephanie Dufek, Adam P. Levine, Mari Thomas, et al. 2022. 'Identification of a Novel Genetic Locus Associated with Immune-Mediated Thrombotic Thrombocytopenic Purpura'. *Haematologica* 107 (3): 574–82. <https://doi.org/10.3324/haematol.2020.274639>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.