IMPLEMENTATION STORIES

FAIRsFAIR
Fostering Fair Data Practices in Europe

THEME 1

ENSURING TRUSTED CURATION OF FAIR DATA

IMPLEMENTATION STORIES

THEME 1 ENSURING TRUSTED CURATION OF FAIR DATA

Making the case for FAIR Data Points

Authors
Angus Whyte, Ryan O'Connor (Digital Curation Centre) and Josefine Nordling (CSC)

From interviews with
Erik Flikkenschild (Leiden University Medical Centre) and Hans van Piggelen (SURF)

Leiden University Medical Center

SURF

Introduction

As a service manager how may I assist my organisation to make research data we hold both FAIR and "as open as possible, as closed as necessary"? The FAIR Data Point is one answer, championed by GO-FAIR as a solution to this need. It offers a metadata publishing component, which is also usable for data visiting, in cases where there are no strong access control requirements (demonstrating). In this story we describe how two organisations have applied the FAIR Data Point (FDP) to provide FAIR data or metadata in two contexts. In Leiden University Medical Centre the FDP is used to publish metadata about COVID patient data (WHO e-Crf) as open as possible in the interest of research, while the data is necessarily closed and held in a variety of different systems. By contrast, Dutch data service provider SURF is applying the FDP to improve the FAIRness of an extensive dataset repository that is openly accessible by default. Based on interviews with the lead protagonists in both organisations' FDP implementations we compare their rationales and approaches, and how they expect this FAIR-enabling technology to benefit their user communities.

FAIRsFAIR recommendation

"Researchers and data stewards should work within GDPR and IPR regulations to make data FAIR and 'as open as possible, as closed as necessary'"

FAIRsFAIR Recommendations on practice to support FAIR principles

Clarifying the FAIR Data Point: two cases

The FDP is a key element of the practical implementation of the vision expressed in the FAIR Principles. The FAIR principles themselves are independent of the technical approach to making datasets (or information about them) as findable, accessible, interoperable and reusable as it can be. They are also independent of the policy decisions and regulations that determine how open it can be. But the premise of this story is that technology designed to be FAIR-enabling from the start can help overcome some of the challenges presented by legacy standards that may be less well-equipped for the task. The FDP uses standards that deal with multiple metadata formats and data sources in a consistent way. While there are other repository platforms that do so, the FDP is specifically designed to make FAIR metadata accessible efficiently, using linked open data standards to expose the semantic relationships in machine-actionable format.

The FDP is a proposed standard for implementing FAIR-enabling repository technologies. It has been developed primarily by the GO-FAIR initiative and the The Dutch Techcentre for Life Sciences (DTL), in conjunction with Leiden University Medical Center (LUMC). FAIRsFAIR also set up a reference implementation, that is described elsewhere[1].

Here we describe approaches, challenges and impacts resulting from two related implementations. The first is at Leiden University Medical Centre, led by Erik Flikkenschild in the context of the Covid-19 research project VODAN, and also involves 84 health and medical centres in ten African countries. The second focuses on Hans van Piggelen's work to improve SURF's open data repository, based on the FDP specifications and DTL reference implementation.

Leiden UMC partnership with VODAN-Africa

Leiden University Medical Centre (LUMC) is the university hospital affiliated with Leiden University in the Netherlands, hosting its medical faculty[2]. LUMC has education as a core mission as well as healthcare and research. As the organisation's information manager for research, and a director of the GO-FAIR Foundation, Erik Flikkenschild with a team of Human genetics bioinformatics experts UL/LIACS and the Go FAIR Foundation has been at the forefront of developing practical approaches to implement the FAIR principles, including the FDP.

The research challenges posed by the COVID-19 pandemic have offered a proving ground, in the shape of the Virus Outbreak Data Network (VODAN) and in particular the VODAN-Africa project. Since 2020 this collaboration of researchers and health practitioners across fifteen African countries has been working to design an inclusive data management system for health in Africa. As Erik explains: "When the COVID project was financed we started to bring the African countries into the lead to demonstrate the value of the FAIR principles. The LUMC was ahead of things, due to the (EU) rare diseases projects and had already set up FAIR data points, so a Task Force was set up to join up these initiatives, towards the goal in Africa which is to change the whole information system into FAIR using open source systems."

African health researchers have historically faced lack of access to local data, due to a variety of factors including lack of ownership of health data, lack of genetic diversity in the data collected, obstacles to digitalisation of health care, monopolistic data practices by non-African research organisations, and diverse policy and legal jurisdictions. The VODAN-Africa project has sought to make a difference by working with African researchers aiming to build an East African Open Science Cloud for Health Data Access.

1. Claudia Behnke, Kees Burger, Yann le Franc, Wim Hugo, Pekka Järveläinen, Jessica Parland-von Essen, & Gerard Coen. (2021). D2.6 First reference implementation of the data repositories features (1.0). Zenodo. https://doi.org/10.5281/zenodo.5362027
2. https://www.lumc.nl/?setlanguage=English&setcountry=en

The latter initiative adopted the FAIR principles in 2018, seeing the opportunity to develop an agile data analytic system, using semantic web standards to work with anonymised patient data, exchanged between clinics and hospitals via mobile phone networks and the Internet. The advent of COVID-19 brought with it a pressing need to efficiently exchange information from clinical reports of disease occurrence[3].

In the interest of shaping responses to the pandemic that would recognise local African needs, while also conforming to global standards, VODAN-Africa began in 2020 with a need to address both the policy (governance and permission) and technical contexts for FAIR implementation (federated patient data analytics) . The project addressed the policy angle by examining the 'FAIR equivalency' of national health regulations and policy documents. This meant comparing (per country) the content of all relevant documents with the 15 FAIR principles. At the same time, the project worked with the World Health Organisation's 'SMART Guidelines' for standardised machine-readable data, and in particular the standard e-CRF (electronic clinical report form)[4].

The idea of 'VODAN in a box' was borne out of the need for a pilot phase to quickly establish a federated network, linking LUMC with servers initially in six African countries (Uganda, Ethiopia, Nigeria, Kenya, Tunisia, and Zimbabwe). Each would need to be able to host FAIR data and support machine access to it, while operating in very different technical and policy contexts. This meant distilling the approach to three key elements, which Erik describes in the context of data orchestration, as 'FAIR data hosting':

1. 'FAIR Implementation Profile' for the community (VODAN in this case) to define its technical (red principles) parameters for implementing FAIR[5]
2. Metadata template representing the e-CRF[6]
3. Interfaces for entering and updating metadata produced using this template, classifying it to semantic metadata standards (DCAT and Dublin Core), registering it using linked open data format (RDF) in a triple store, and for supporting queries using a SPARQL endpoint[7].

The third of these elements corresponds to the FDP specification, and can be broken down into three technical components - an API, a service to implement this API and support metadata definition, and a client of the API to handle updates and queries (for more information see[8]).

It was vital for the VODAN approach to implementing the FDP to support the concept of 'data visiting'. Making data 'FAIR for machines' is one of the basic tenets of GO-FAIR's vision for FAIR, and 'data visiting' is the means by which machine-actionable data can transform data reuse. Trust in data ownership, data locality and GDPD responsibility were the key factors LUMC has played a key role in engineering and demonstrating a practical approach to this, as part of the African medical community's battle to monitor the

3. van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihgo, M., Folorunso, S., Schultes, E., Basajja, M., Aktau, A., Amare, S. Y., Taye, G. T., Purnama Jati, P. H., Chindoza, K., Wirtz, M., Ghardallou, M., van Stam, G., Ayele, W., Nalugala, R., Abdullahi, I., Osigwe, O., Graybeal, J., ... Musen, M. A. (2021). Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced genetics (Hoboken, N.J.), 2(2), e10050. https://doi.org/10.1002/ggn2.10050
4. Mehl, G., Tunçalp, Ö., Ratanaprayul, N., Tamrat, T., Barreix, M., Lowrance, D., Bartolomeos, K., Say, L., Kostanjsek, N., Jakob, R., Grove, J., Mariano, B., Jr, & Swaminathan, S. (2021). WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age. The Lancet. Digital health, 3(4), e213–e216. https://doi.org/10.1016/S2589-7500(21)00038-8
5. VODAN FIP available at: https://osf.io/gz2st/
6. CEDAR revised case report form for Confirmed Novel Coronavirus COVID‐19 https://openview.metadatacenter.org/templates/https://repo.metadatacenter.org/templates/3dd62fd4-3e77-414f-b5f5-01d25acf54f6
7. Suchánek, Marek, & Mariam Basajja. (2020). VODAN in a Box: Proof of Concept. Zenodo. https://doi.org/10.5281/zenodo.4321626
8. https://www.fairdatapoint.org/

COVID-19 outbreaks in their countries. As Erik Flikkenschild summarises the rationale:

> *"WHO asked the world community to start with one e-CRF that could be used worldwide, and with a data visiting approach you could count a lot of things in a more efficient and secure way like the number of patients, and monitor if things are getting worse."*

Data visiting refers to two concepts, distributed learning, and federated learning, The (Health-RI) Personal Health Train concept concept addresses this. VODAN uses a Virtual Research Environment to provide a trusted community workspace with a FDP or 'FAIR station', which is based in (e.g) a hospital and can be queried and analysed remotely to support research decisions. Figure 1 gives a high-level overview of the approach.

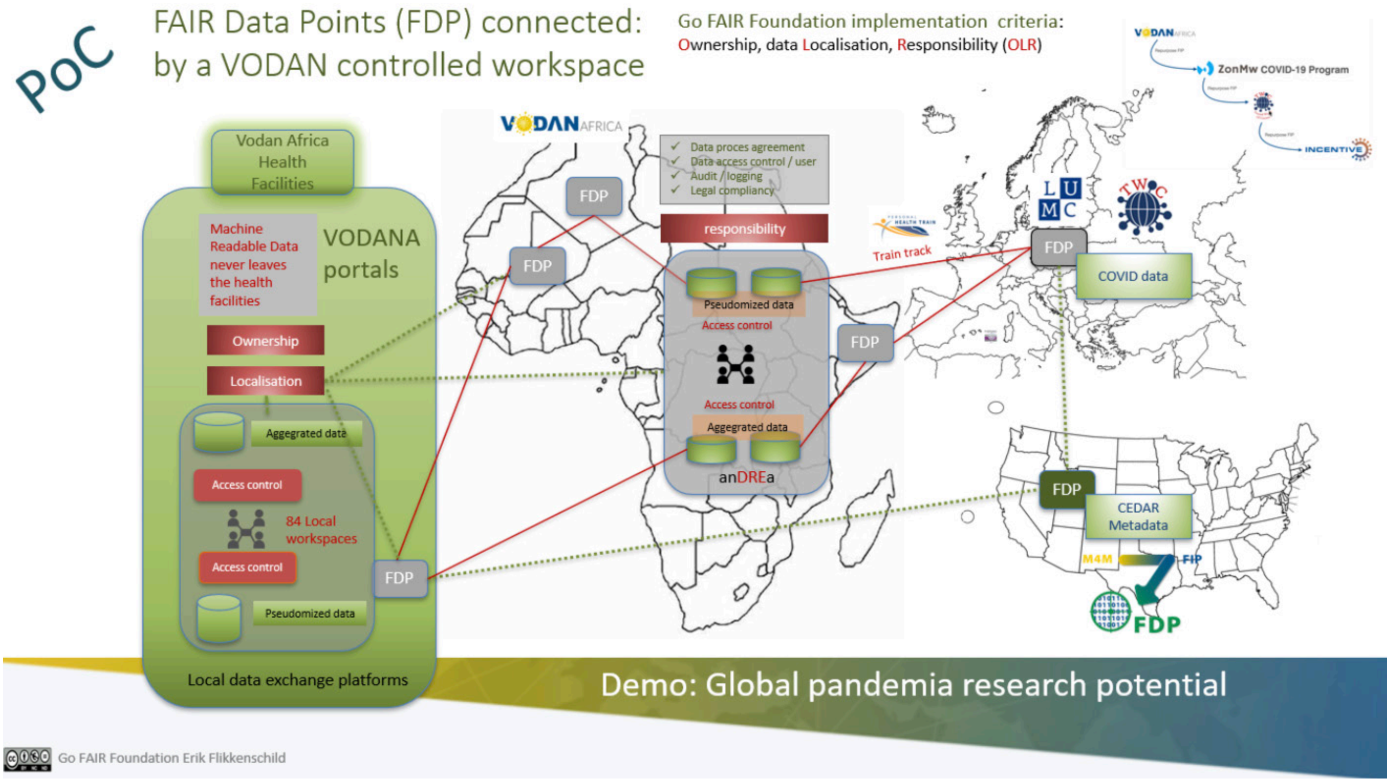


Figure 1. VODAN Governance view of data visiting[9]

The concept of algorithms 'visiting' FAIR Data stations based on the FDP architecture, to conduct analysis there, is key to the federated approach VODAN takes to FAIRifying patient data. The approach is not purely technical in concept, as it also involves supporting local data governance to ensure trustworthy curation of the data by local data stewards, while working within disparate regulatory jurisdictions.

9. https://osf.io/98uvp/

Challenges encountered and addressed

The data visiting scenario requires various social and organisational arrangements to be in place, as well as technical components. The VODAN-Africa initiative faced considerable policy-related, social, and organisational challenges, and addressing these has helped shape GO-FAIR's thinking about the alignment of technology, policy, and research data practices.

As a result, Erik believes, "We have learned how to organise a FAIR competence centre effectively. The main challenge is for everyone to really understand the FAIR principles, and it can be an effort to do this. Some claim they are aware of what the implications of the FAIR principles are, however in reality, that might not be the case." He relates this challenge to LUMC's educational role, and has addressed it by developing 'train-the-trainer' resources.

Data stewards play a focal role in a FAIR competence centre, and this focus needs to engage policy-makers and technologists as well as researchers producing and reusing data. A starting point is to define the explicit role of policy makers. "This stakeholder group is also a challenge", Erik acknowledges, " as one needs to convey the right message in order to make them understand on a sociological level what this is all about."

"The case for an FDP in a healthcare context needs to be developed with a privacy officer, as you have to build the trust in using the technology, and need to solve the ethical issues to be able to use it in a way that addresses the risks. GDPR adds a true challenge here, however, the African countries decided to adopt the GDPR to have consistent policy and access control for patient information to address privacy concerns, and to respect the ownership of the data."

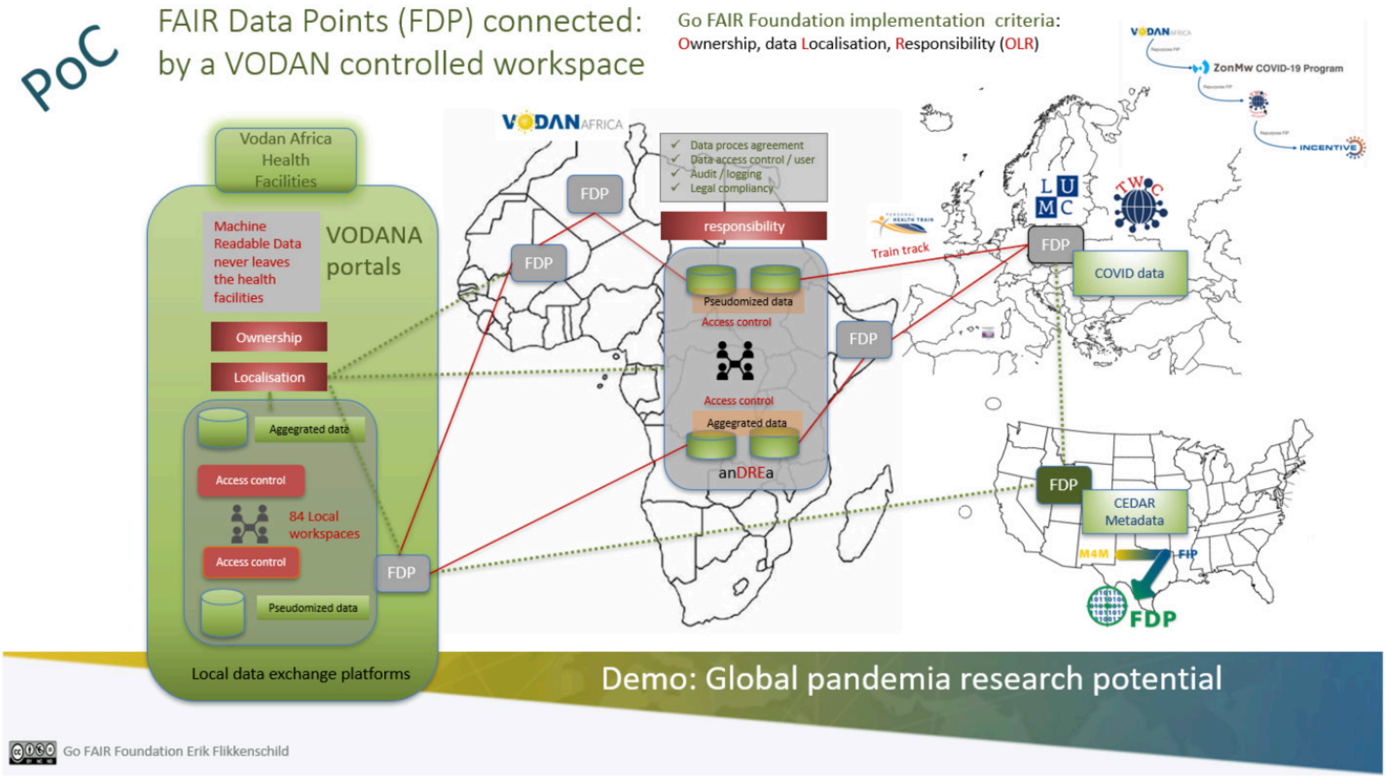
The main consideration is that benefits of data reuse need to be evident, to convince the various stakeholders - policy, IT experts and higher management - that it is possible to address concerns for the institutional community.

While privacy and ethical concerns are not new, data visiting adds a novel twist to them, and makes it more essential to have formal processes in place before an FDP can support this automated form of data reuse. "To connect with real patient data you have to build access control policies on the ports, and you have to define the appointment system." This determines when an algorithm can visit the metadata, and defines the agreements between the organisations involved: the processor and controller agreement (DPA), the data licence, and method for gaining informed consent.

Impacts - a proof of concept in VODAN-Africa

Erik Flikkenschild sees the 2021-22 period as a turning point for organisations to see the advantage of having reusable data, and points to the success of VODAN-Africa in demonstrating 'proof of concept' of data visiting. Between July and September 2020, a total of 10 FAIR Data Points were installed across the countries involved in the project, each coupled with the electronic WHO COVID report form (e-CRF).

Once these FDPs were online, a proof of concept test was carried out to run queries between LUMC and the FAIR Data Point of Kampala International University. This was successfully realized, when queries were run across the two continents in September 2020 and produced aggregate findings from analysis of patient records. While this was limited to the relatively limited scope of the e-CRF, further work is needed to support analysis from FDPs that integrate both clinical and research data and the workflows for their production (van Reisen et al 2021 gives further information on this).



9. https://osf.io/98uvp/

Improving SURF's Data Repository

Aims and approach

SURF is a collaborative organisation for IT in Dutch education and research. It recently implemented a FAIR Data Point to realise an ambition for the SURF Data Repository, which serves researchers across the Netherlands and users of their data internationally.

The SURF Data Repository was fully established in 2021, specifically targeting the publication of large datasets (upwards of 3 TB) which need to be stored and made accessible in a cost-effective way. It offers users the ability to "discover and safely store, annotate and publish research data of any scientific domain".

SURF already offered a data archive, but the data and metadata stored there were mostly not FAIR due to the closed access nature of that service, according to SURF consultant Hans van Piggelen. Hans and his colleagues saw the need to change this, and wanted to develop a production-level implementation of the FDP specification, inspired by a GO-FAIR hackathon at the Dutch Technocentre for Life Sciences (DTL).

The repository is built upon Fedora Commons. This provides the underlying software layer for managing digital objects, and includes a triplestore for registering semantic relationships between these objects, using RDF. To manage and publish the metadata SURF turned to the FDP. As Hans puts it:

> *"The FDP seems more versatile than other platforms- in that it is better at representing structure - one can more easily go into setting up a recursive structure for the data which is needed for our collections - and not all other repository platforms are good at handling the concept of collections. The FDP allows the repository to handle an infinite level of hierarchy in these collections."*

As well as handling the internal repository content structures more effectively, SURF were looking for better metadata exposure. The repository was previously using an OAI-PMH endpoint, supporting that relatively older XML-based protocol. "OAI-PMH can easily communicate metadata of digital objects, but it is not really designed for representing and communicating more complex underlying relationships between internal and external objects" says Piggelen.
The FDP improves the ability to find the metadata and see a description, via a SPARQL endpoint. "That will not necessarily overcome any barrier to accessing the material itself" says Hans, "The FDP can be configured to apply different access policies, but for the SURF repository that is not a problem as it offers access which is open by default".
SURF also wanted to make it possible for algorithms to visit the data, and valued the standards-based approach to this adopted in the FDP. The SURF repository has not supported this explicitly yet, but expects to do so in future.

10. https://www.surf.nl/en
11. https://repository.surfsara.nl

Challenges encountered and addressed

The familiarity with the FDP that was gained in the DTL hackathons made the implementation less effort than it might have otherwise been. It offered an opportunity to roadtest some features relevant to their aims. The 'open by default' policy adopted for the SURF repository also made access control less challenging. Access policies applicable to a dataset would be configurable by data owners depositing the datasets, a role that could involve a data librarian or a data steward.

One initial challenge for the SURF team was to map the available classes and properties of the DCAT vocabulary used in FDP with the terms and structure used with the data published in the repository. "This worked out quickly towards full exposure" says Hans, although "the hierarchical nature made this more complex initially".

The main technical challenge was the need to know how to convert the metadata held by the repository into the SPARQL endpoint. This requires technical expertise and knowledge of the metadata, how it is stored and how it is structured internally. In Hans' estimation, if repository implementers are equipped with this structural information about metadata structures it should generally be feasible to set up an FDP within a month on top of existing repository infrastructure and services. If a completely new FDP service is set up, this could take longer, depending on the required additional service components that need to be set up.
The generic nature of the repository and lightweight approach to curation also made the FDP implementation relatively straightforward. A domain repository would be likely to have more complex access policies and licence conditions, according to Hans Piggelen. The SURF repository does not set rules for the deposit of data in discipline-specific formats. Instead it supports 'communities' that are, in effect, bundles of datasets accompanied by community-specific metadata defined in additional metadata schemas, and managed by an administrator with the appropriate privileges. Communities can also have user groups, to provide access or admin privileges to specific individuals. However SURF does not get involved in this level of curation. As Hans explains, "we don't look at the data contents...we would make changes if needed for a format conversion, but only on request of the data owner. Images might be convertible in that sense, but we would not do any interpretation of the actual contents. If data curation is needed, it is considered on a case-by-case basis and again only on request by the data owner."

Impacts - providing a Dutch service for large datasets

Having passed the milestone of technical implementation the SURF Data Repository went live in its new FDP-based incarnation in 2021. It has been adopted since then for EOSC-related purposes. SURF is a partner in DICE (Data Infrastructure Capacity for EOSC), a Horizon 2020 project. In that context, the repository now hosts large-scale Astronomy datasets from a sky survey conducted by LOFAR, a large radio telescope network located mainly in the Netherlands, as well as several datasets from other Dutch research institutes

More generally, the repository has not been available for long enough to identify impacts, in terms of concrete differences for the research communities that it targets. However SURF fully expects it to add value for these communities. Hans van Piggelen identifies a number of impacts he expects from better enabling FAIRness.
Findability is the most straightforward benefit. In Hans's view, improving the discovery of data stored in the repository is the most important end goal that the FDP satisfies, for example by using RDF triples to improve representation of hierarchical relationships in the datasets.

He also sees very interesting opportunities for working in a decentralised way to conduct data analysis, "it makes it possible to use as much data as there is, without having to transfer or copy data" he observes, "... and distributed search could be very useful, and the possibility to run algorithms without the second party being able to read the data directly, and without any data to be transferred.

Van Piggelen also sees a great opportunity for technical progress in the development of micro-services that can operate across distributed networks of FDPs.

Conclusions

The FDP is essentially a standardised approach to providing more semantically rich and machine-actionable metadata, whether 'out of the box' or as a layer on an existing repository platform capable of delivering semantic data. The two stories of FDP implementation we have described are from very different contexts, each showing that, with the necessary competences, organisational arrangements and policy contexts in place, deploying an FDP can make a step change in data reusability and demonstrate the value of FAIR data.

Further information

VODAN-Africa

van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihgo, M., Folorunso, S., Schultes, E., Basajja, M., Aktau, A., Amare, S. Y., Taye, G. T., Purnama Jati, P. H., Chindoza, K., Wirtz, M., Ghardallou, M., van Stam, G., Ayele, W., Nalugala, R., Abdullahi, I., Osigwe, O., Graybeal, J., … Musen, M. A. (2021). Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced genetics (Hoboken, N.J.), 2(2), e10050. https://doi.org/10.1002/ggn2.10050

SURF Data Repository

SURF Data Repository https://repository.surfsara.nl

Research Data Netherlands 'Research Data Decision Tree'
https://researchdata.nl/research-data-decision-tree/

T. W. Shimwell et al. (2022). The LOFAR Two-metre Sky Survey (LoTSS) V. Second data release (data set), https://doi.org/10.25606/SURF.lotss-dr2

About FAIRsFAIR Implementation Stories

FAIRsFAIR Implementation stories illustrate good practices in research communities and organisations to support the implementation of the FAIR principles. These practices encompass 'FAIR-enabling' actions as recommended in the EC Expert Group on FAIR report Turning FAIR into Reality and the FAIRsFAIR Recommendations on practice to support FAIR principles. FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558. The content of this document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of such content.

FAIRsFAIR - Fostering FAIR Data Practices in Europe

European Commission Grant Agreement No. 831558
H2020-INFRAEOSC-2018-4
www.fairsfair.eu - support@fairsfair.eu

Acknowledgements

This report has been produced by the FAIRsFAIR (GA No. 831558) project, which received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-05-2018-2019. FAIRsFAIR gratefully acknowledges everyone who we could interview for this series.

Disclaimer

Copyright rests with the authors