FAIR Bioinformatics computation and data management: FAIRDOM and the Norwegian Digital Life initiative.

Natalie J Stanford[1], Finn Bacall[1], Fatemeh Zamanzad Ghavidel[2], Martin Golebiewski[3], Inge Jonassen[4], Rune Kleppe[2], Olga Krebs[3], Hadas Leonov[3], Stuart Owen[1], Kjell Petersen[4] Maja Rey[3], Stian Soiland-Reyes[1], Kidane Tekle[4], Andreas Weidemann[3], Alan Williams[1], Ulrike Wittig[3], Katy Wolstencroft[5], Anders Goksøyr[6], Jacky L. Snoep[1,7], Jon Olav Vik[8], Wolfgang Müller[3] and Carole Goble[1]

1 School of Computer Science, The University of Manchester, United Kingdom;
2 Centre for Digital Life Norway, University of Bergen, Norway;
3 Heidelberg Institute for Theoretical Studies, Germany;
4 ELIXIR Norway, University of Bergen, Norway;
5 Leiden Institute of Advanced Computer Science, Netherlands;
6 Department of Biological Sciences, University of Bergen, Norway;
7 Stellenbosch University, South Africa;
8 Department of Animal and Aquacultural Sciences, Faculty of Life Sciences, Norwegian University of Life Sciences, Norway

Corresponding author: natalie.stanford@manchester.ac.uk

Introduction

The FAIR data principles state [1] that data should be Findable (and citable), Accessible (with appropriate caveats for sensitive data), Interoperable (able to be exchanged or combined, typically as a result of using community standards) and Reusable (can be reused later or reproduced from a publication). The FAIRDOM Research Infrastructure (http://fair-dom.org) has been fostering software and community activities to support the adoption of FAIR data management practices since 2008. It supports research projects with their research data, methods and model management, at all stages, emphasizing standards and sensitivity to asset sharing and credit anxiety, and providing consultancy and training services. While conceived with a focus on Systems Biology, FAIRDOM's infrastructure is domain agnostic and has been adopted and adapted to support, for example, biodiversity simulation workflows. Nevertheless, its prime adoption community is Life Sciences.

The FAIRDOM Platform supports a Commons that brings together the outcomes of projects, researchers and their investigations, linked with resulting publications or presentations. The SEEK software component of FAIRDOM [2] provides a web-based yellow pages (of programmes, projects, institutions and people), a catalogue of their research assets (data, models, SOPs, samples, organisms, publications etc), and the metadata associated with them. A common, related view of the assets is organised using ISA (Investigation, Study, Assay/Analysis).

FAIRDOM is very flexible regarding the location of content, which can be uploaded to the FAIRDOM databases for small datasets or held externally in third party databases where it can be referenced. Thus the catalogue can span many different datastores including: e-infrastructures installed "on premise" such as the openBIS management and analysis platform [3] (an optional back-end of the FAIRDOM Platform); e-infrastructures provided nationally such as ELIXIR-Norway's National e-Infrastructure for Life Sciences (NeLS); and public archives such as the Core Data Resources and Recommended Deposition Databases of ELIXIR. Recorded outcomes are typically held in local stores or deposited in type-specific specialized public repositories and catalogues. Scattering results across repository silos loses the vital gathering and contextualization of data and models key to integrative biology research. FAIRDOM overcomes this to pool results and link public resources with project data whilst retaining a structured synthesis of the data, SOPs and models associated with different investigations.

To date:

- The FAIRDOM Platform has been deployed by over 30 projects and centres to support their asset management, including two Synthetic Biology Centres in the UK;
- The FAIRDOMHub Commons [4] public resource allows projects to self-manage their results, and manages the outcomes of over 100 projects, with over 1000 registered researchers from more than 220 institutions.

Over 35 national and pan-national programmes use the FAIRDOMHub for data retention and sharing of their projects, including ERA-NETs (e.g. SysMO, ERASysAPP), the German Network for Bioinformatics Infrastructure (de.NBI) and the Centre for Digital Life Norway.

The Centre for Digital Life Norway (DLN) is a virtual centre funded by the Research Council of Norway under the Digital Life initiative. It supports and promotes the use of data management for its projects, including the Digital Salmon (DigiSal) and dCOD 1.0. DigiSal is assembling a library of life process models in the salmon body to construct tailored simulations, whereas dCOD 1.0 aims at decoding the systems toxicology of atlantic cod. Both are typical Systems Biology projects:

- multi-disciplinary, with a range of research outputs (models, data, samples, Standard Operating Procedures, publications) that are interrelated, versioned and shared to varying degrees internally and externally to the projects at different points in the research lifecycle.
- multi-partner, with collaborations across national and international institutions.

It is in the interests of the projects that their data is FAIRly stewarded for effective collaboration, analysis and ultimately for publishing. It is in the interests of the centre that project outcomes are retained and FAIR for the benefit of other projects, and for the demonstration of investments by the funders.

The Norwegian e-Infrastructure for Life Sciences (NeLS) is developed and supported by ELIXIR-Norway as a national infrastructure for bioinformatics, providing storage, data sharing and analysis tools (using Galaxy) and connecting to the national data storage platform, NIRD, for long term storage. Thus NeLS is a resource to enable bioinformatics researchers to flexibly store and compute data, from raw to derived, using bioinformatics workflows that allow high throughput raw data to be stored temporarily while it is transformed through a series of workflows to processed derived data.
FAIRDOM, DLN and ELIXIR-Norway teamed up to support projects in the Digital Life initiative with their FAIR data management.

Methods

FAIRDOM offers rich metadata and enables researchers to share their data with all partners in context. FAIRDOMHub had been adopted by the Digital Salmon and dCOD for this purpose (https://fairdomhub.org/programmes/7 and https://fairdomhub.org/programmes/22 respectively). For example, in Digital Salmon, 38 researchers from 7 different institutes can privately find, access and reuse datafiles and associated SOPs from 10 investigations, 20 studies and 65 assays. Once the research has been published, these can be made publicly accessible and a snapshot or results assigned a DOI through a simple publication process.

NeLS provides nationally supported data storage, but the raw and derived data needs to be annotated and catalogued in context with other data and models from the research project or publications, in order to adhere to the FAIR principles.

Moreover, NeLS is only accessible to researchers within Norwegian institutes and their close international collaborators in a restrictive manner, whereas projects need more flexible access for other project members beyond NeLS. By making the NeLS platform compatible with the SEEK platform, DLN researchers can share their data in context after they have computed their results using the NeLS analysis tools (Figure 1).

Linking a NeLS file and metadata to SEEK

- A user of a SEEK project creates a reference to a data resource in NeLS by opening a dialogue to log on to NeLS, navigating through the NeLS portal to identify and select a desired processing level (sub-datatype) of a data set, and then referencing this in SEEK as an Asset or DataFile. Metadata about Samples and other relevant dataset entities can be stored together with the data in NeLS, and be imported by SEEK as part of the process creating the reference. Metadata can be stored using semantically annotated spreadsheets from the software RightField [5], formatted compatible for NeLS, so that it can travel with the data into SEEK.

Accessing NeLS from SEEK

- Any user, within a project where the data file is visible in FAIRDOMHub can see the NeLS reference and open the link to view the data in NeLS, subject to having the correct authentication. The user can thereafter continue working within the NeLS portal on computation and/or management of that data set.

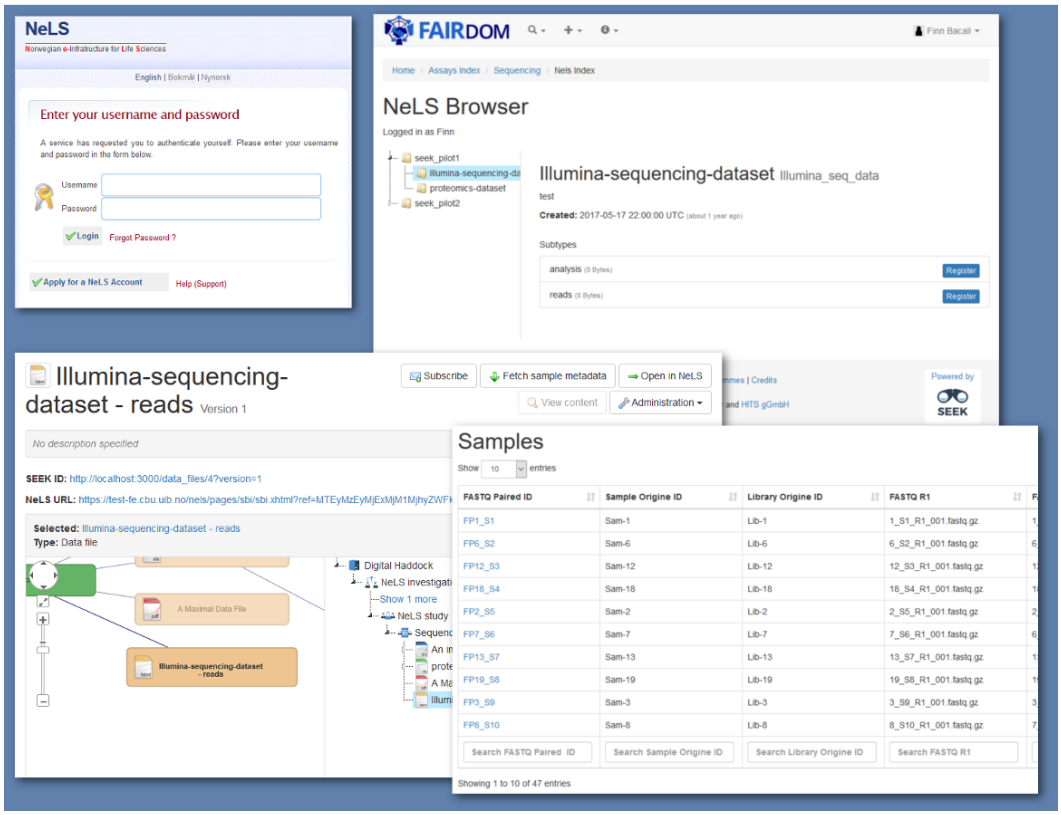

Figure 1: Showing how data is imported into SEEK from the NeLS portal. Users who have access to NeLS are able browse datasets in the NeLS Browser. Once a suitable dataset is identified it can be imported into SEEK as Samples. From here the samples can be shared with a broader audience, be assigned DOIs, and linked to within publications.

Results

The Summer 2018 releases of NeLS and SEEK allow users with accounts in NeLS and FAIRDOMHub to place their bioinformatics data in a temporary storage and compute environment, and transfer the data into long-term storage for publication. This improves FAIR support for bioinformatics data in the following ways:

- *Findable*: Researchers can record their computed data in context with other data from their projects, within the SEEK. This is facilitated through a simple import interface. This recording of data in SEEK makes it more Findable for other project members, and after publication. DOIs can be assigned to the data, to be used as shareable links in publications.
- *Accessible*: Researchers registered with NeLS are able to follow the link from SEEK through to NeLS and download, or perform further computation/analysis on the data if they have the appropriate permissions. After publication, a broader audience is able to access the data and associated metadata through SEEK.
- *Interoperable*: Researchers are able to import data files from NeLS that are then converted into individual samples in SEEK using a compatible template.
- *Reusable*: The inclusion of metadata annotation with the data ensures that the data is fully described and therefore easier to re-use.

The interface has been tested by a number of NeLS and SEEK programmes including Digital Salmon and dCod 1.0, and is now in active use for computation, and management of their bioinformatics data.

Conclusions

The NeLS-SEEK integration tackles the so called “last mile” issue - to bridge between researchers’ (meta)data management needs and the facilities offered by national research e-infrastructures [6]. In order to make the FAIR principles enactable over the whole data management pipeline, integrations such as the NeLS-SEEK integration will need to become commonplace in data management of research projects. Integration of key resources should allow seamless exchange of data through all the processes of a data management pipeline. Removing the manual element of data exchange improves the likelihood that researchers will transfer their data through the pipeline, and maintain the FAIR data as they move data from machine collection through to publication.

Overall FAIRDOM, and specifically the integration of its SEEK catalogue component with NeLS, supports the pathway that links researchers day-to-day, and provides a way to deposit their data with minimum manual effort into research e-infrastructures, and to link public content with their own.

References

1. Wilkinson M, Dumontier M et al (2016) The FAIR Guiding Principles for scientific data management and stewardship Scientific Data 3, doi:10.1038/sdata.2016.18
2. Wolstencroft K, Owen S, Kreb O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Muelle W, Goble C (2015) SEEK: a systems biology data and model management platform", BMC Systems Biology, 9(1):1-12 doi:10.1186/s12918-015-0174-y
3. Bauch A, Adamczyk I, Buczek P, Elmer F-J, Enimanev K, Glyzewski P, Kohler M, Pylak T, Quandt A, Ramakrishnan C, Beisel C, Malmström L, Aebersold R and Rinn

B (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research, BMC Bioinformatics 12(468) doi:10.1186/1471-2105-12-468
4. Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, Kuzyakiv R, Nguyen Q, Owen S, Soiland-Reyes S, Straszewski J, van Niekerk DD, Williams AR, Malmström L, Rinn B, Müller W, and Goble C (2016) FAIRDOMHub: a repository and collaboration environment for sharing systems biology research, *Nucleic Acids Research*, 45(D1) pp: D404–D407 doi:10.1093/nar/gkw1032
5. Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, du Preez F, Goble CA (2011) RightField: Embedding ontology annotation in spreadsheets. Bioinformatics 15;27(14):pp2021-2 DOI: 10.1093/bioinformatics/btr312
6. Koureas D, Arvanitidis C, Belbin L, Berendsohn W, Damgaard C, Groom Q, Güntsch A, Hagedorn G, Hardisty A, Hobern D, Marcer A, Mietchen D, Morse D, Obst M, Penev L, Pettersson L, Sierra S, Smith V, Vos R (2016) Community engagement: The 'last mile' challenge for European research e-infrastructures. Research Ideas and Outcomes 2: e9933. https://doi.org/10.3897/rio.2.e9933

Funding Acknowledgements

FAIRDOM received funding support for this work from the UK Biotechnology and Biological Sciences Research Council BB/M013189/1 (DMMCore). German Federal Ministry of Education and Research (BMBF) as part of the German Bioinformatics Infrastructure (de.NBI: FKZ 031A371), the German Liver Systems Medicine Network (LiSyM: FKZ 031L0056) and the European ERA-Net for Systems Biology Applications (ERASySAPP: FKZ 031A525), as well as from the Klaus Tschira Foundation (KTS). Research Council of Norway grant nb 248810.