RO-Crates as a practical implementation of FAIR Digital Object to align biodiversity genomics work streams

RO-Crates as a practical implementation of FAIR Digital Object to align biodiversity genomics work streams

Authors:

- Tom Brown (Leibniz Institute for Zoo & Wildlife Research, ERGA, BGE)
- Nick Juty (The University of Manchester, Bioschemas, ELIXIR-UK, FAIRplus, BY-COVID, FAIR-IMPACT, BGE)
- Stian Soiland-Reyes (The University of Manchester, University of Amsterdam, RO-Crate lead, ELIXIR-UK, BY-COVID, FAIR-IMPACT, EOSC-Life, EuroScienceGateway, BGE)
- Felix Shaw (Earlham Institute, BGE)
- Rutger Vos (Leiden University, Naturalis Biodiversity Center, BGE)

Category:

- Science
- Technology
- FAIR

Keywords: FDO, Data management, metadata, biodiversity

Summary

We describe our pragmatic approach for aligning parallel scientific processes through the implementation of Fair Digital Objects (FDOs), as RO-crates. Our work is grounded in the Biodiversity domains, but may be extrapolated to be useful more generally in other scientific domains.

Proposed topic (2 pages)

ERGA stream

The European Reference Genome Atlas (ERGA) is a pan-European scientific response to threats to biodiversity, with the aim of producing reference genomes for all eukaryotic species in Europe via a distributed model of sampling, sequencing, and generating genome assemblies. This process involves a number of distinct sets of data, metadata, and workflows, which must be linked and published in a connected manner. Existing infrastructure such as provided by the INSDC (ENA, BioSamples, GenBank, DDBJ [Arita et al., 2021]), COPO [Shaw et al., 2020], and WorkflowHub [Goble et al., 2023] are currently used to collect raw sequencing data, sample metadata, genome assemblies, and workflows, respectively, yet the nature and clarity of their inter-connectedness remains a challenge. For research outputs of ERGA to remain FAIR [Wilkinson et al., 2016], the genomics field requires the development of resources such as RO-crate [Soiland-Reyes 2023], where the provenance and interconnectedness of research objects can be explicitly stated and maintained throughout the life of a genome assembly project. Supporting RO-crate for submission to public nucleotide archives such as ENA or NCBI would facilitate the attachment of associated metadata, data, and workflows to research outputs, and ensure the relations between these objects are explicit.

iBOL stream

BIOSCAN is the second tranche of bundled activity within the DNA barcoding community, operating under the aegis of the iBOL consortium, to achieve a step change in the number of publicly available, high-quality reference DNA barcode sequences used for species identification, delimitation, and systematics. DNA barcoding refers to the concept of sequencing single-copy, single-inheritance, genetic markers whose constant substitution rate is such that they have low variation within species but sufficient variation between species to allow for discrimination among them [Hebert et al., 2003]. The BIOSCAN ‘stream’ (slated to be renamed iBOL Europe) is the practical implementation of this step change within the context of the Biodiversity Genomics Europe project. The implementation involves a number of work processes starting with specimen collection (both in the wild and from natural history collection), sampling, DNA extraction, library preparation, sequencing, bioinformatics, quality assurance, and data publishing to BOLD [Ratnasingham & Hebert, 2007], the central repository for DNA barcodes. Due to the variation in the sample sources - fresh or preserved - the chain includes some branching depending on whether the DNA extracts are amplicon sequenced or genome skimmed, with downstream impacts on bioinformatics and QA. The complexity of the processes and their distributed nature complicate provenance tracking and FAIRness, which is nevertheless essential for the credibility of the resulting data products. To address this, the BIOSCAN stream adopts RO-crate technology as the common framework.

RO-Crate

RO-Crate [Soiland-Reyes 2023] is a community effort to practically achieve FAIR packaging of research objects (digital objects like data, methods, and software) with structured metadata. RO-Crate uses well-established web standards and FAIR principles. For common metadata representations, RO-Crate builds on schema.org, a mature and general mark-up vocabulary used by search engines including Google Dataset Search. RO-Crate is adapted by many EU/EOSC projects as a

pragmatic implementation of the FAIR Digital Objects [De Smedt et al., 2020] vision. This paper brings together representatives of existing Biodiversity ‘pipelines’ in an effort to identify and harmonise processes and (meta)data that are currently parallel but equivalent. We envisage convergence on a common, or at least more integrated, scientific and computational workflow where (meta)data are better interrelated both within a workstream and across them. The mechanism by which we will achieve this is the RO-Crate implementation of FDOs. This work is relevant to biodiversity projects and community-driven efforts but also serves as a practical solution to comparable issues which will exist in other domains, by analogy.

Implementation strategy

Faced with the opportunities provided by the RO-crate framework, both ERGA and BIOSCAN are developing approaches towards its adoption and implementation. Along the respective streams, various technologies can be leveraged to achieve this. In some parts of the streams, this may require more custom development. For example, both streams struggle with tracking physical specimens and samples as they are being collected, lab-processed, and sequenced. The metadata about the physical objects and their status can be expressed and negotiated among participants in the stream using simple spreadsheets, but the problems these pose in terms of inadvertent data modification (e.g. date parsing) and vague semantics may drive the development of FDO solutions. Further down the stream, technological developments make RO-crate adoption easier. Once sequenced, the data from both streams is processed by various analysis pipelines. Some of the workflow management systems in which these pipelines are implemented provide RO-crate functionality ‘out of the box’, for example, Galaxy [Galaxy 2022] or NextFlow [Nextflow 2017], while others (such as SnakeMake [Snakemake 2021]) have less well-developed support, which influences workflow management technology choices. How tracking of physical objects and pipeline provenance as FDO is, in the end, integrated and attached to data submissions is likewise a topic that is still being explored by the consortium.

References

[Arita et al., 2021] Masanori Arita, Ilene Karsch-Mizrachi, Guy Cochrane (2021):
The international nucleotide sequence database collaboration.
Nucleic Acids Research **49**(D1)
https://doi.org/10.1093/nar/gkaa967

[De Smedt et al., 2020] Koenraad De Smedt, Dimitris Koureas, Peter Wittenburg (2020):
FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units.
Publications **8**(2)
https://doi.org/10.3390/publications8020021

[Goble 2023] Carole Goble, Finn Bacall, Stian Soiland-Reyes, Stuart Owen, Ignacio Eguinoa, Bert Droesbeke, Hervé Ménager, Laura Rodriguez-Navas, José M. Fernández, Björn Grüning, Simone Leo, Luca Pireddu, Michael Crusoe, Johan Gustafsson, Salvador Capella-Gutierrez, Frederik Coppens (2023): **EOSC-Life Workflow Collaboratory for the Life Sciences**.
Proceedings of the Conference on Research Data Infrastructure **1**
https://doi.org/10.52825/cordi.v1i.352

[Hebert et al., 2003] Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard (2003):
Biological identifications through DNA barcodes.
Proceedings of the Royal Society B **270**(1512)
https://doi.org/10.1098/rspb.2002.2218

[Ratnasingham & Hebert, 2007] Sujeevan Ratnasingham, Paul D. N. Hebert (2007):
BOLD: The Barcode of Life Data System (http://www.barcodinglife.org)
Molecular Ecology Notes **7**(3)
https://doi.org/10.1111/j.1471-8286.2007.01678.x

[Shaw et al., 2020] Felix Shaw, Anthony Etuk, Alice Minotto, Alejandra Gonzalez-Beltran, David Johnson, Phillipe Rocca-Serra, Marie-Angélique Laporte, Elizabeth Arnaud, Medha Devare, Paul J. Kersey, Susanna-Assunta Sansone, Robert P. Davey (2020):
COPO: a metadata platform for brokering FAIR data in the life sciences.
F1000 research **9**(495)
https://doi.org/10.12688/f1000research.23889.1

[Soiland-Reyes 2023] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022):
Packaging research artefacts with RO-Crate.
Data Science **5**(2)
https://doi.org/10.3233/DS-210053

[Wilkinson et al., 2016] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray,

Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons (2016):
The FAIR Guiding Principles for scientific data management and stewardship.
Scientific Data **3**
https://doi.org/10.1038/sdata.2016.18

[Galaxy 2022] The Galaxy Community.
The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update
Nucleic Acids Research, **50**(W1)
https://doi.org/10.1093/nar/gkac247

[Nextflow 2017] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017):
Nextflow enables reproducible computational workflows.
Nature Biotechnology, **35**(4)
https://doi.org/10.1038/nbt.3820

[Snakemake 2021] Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J.
Sustainable data analysis with Snakemake.
F1000Research 2021, **10**:33
https://doi.org/10.12688/f1000research.29032.2