Appendix 4: data management and sharing plan

Overview. This data management plan describes how research data and other research outputs produced at the Quantum Biology Institute (QBI) are created, organized, shared, and preserved. QBI's approach is built around continuous, FAIR (Findable, Accessible, Interoperable, Reusable) [1] publication of research outputs of all kinds, including code, data, protocols, figures, and written materials. **Open and pro-active release of finished research outputs is the default mode of dissemination**. Here, we describe QBI's day-to-day research data workflow, the licenses under which each category of output is released, the intellectual property (IP) consequences of those licenses, and the technical mechanisms used to ensure that outputs are durably identified, protected against loss or tampering, and openly available for reuse. QBI's practices are designed to satisfy the data management and open science requirements of research funders committed to open science and to align with the broader FAIR principles endorsed by funding bodies worldwide.

Roles and responsibilities. The Chief Science Officer at QBI has overall responsibility for the implementation of this data management plan and for ensuring that research outputs are released in accordance with applicable funder policies and QBI's own open science commitments. The Head of Computational Research and Data Infrastructure at QBI is responsible for operating the underlying infrastructure described below, including the QBI electronic lab notebook (ELN) publishing pipeline, storage and backup systems, authentication and remote access, and integration with external archiving venues. Scientists and trainees are responsible for recording experimental and computational work in their ELNs, annotating outputs with appropriate metadata, and depositing finished outputs in the venues described below when or before those outputs are referenced externally or shared with a funder in a written report.

Research output workflow. QBI's research output workflow is designed around the principle that outputs are published continuously as they become ready for reuse, rather than bundled into a later journal article. Day-to-day scientific work is electronically recorded in Jupyter [2]-based lab notebooks authored locally by QBI members. ELNs, source code, and associated data are stored in the appropriate QBI 'vault' on the Frida research server (a RAID10 Dell workstation administered by QBI), which serves as the working environment for the Institute. Access to the server is provided over WireGuard VPN, and shared files are exposed via authenticated Samba shares. All working content on Frida is automatically backed up to Backblaze B2 every six hours. Finished or reusable outputs are promoted from the working environment into QBI's MyST-based ELN publishing pipeline, which renders notebooks and supporting content as a FAIR-native website hosted at the website https://data.quantumbiology.org. This site is organized into vaults that correspond to projects and working groups; each rendered output carries structured metadata and a stable URL. Outputs intended for broader external dissemination, including completed datasets, software releases, protocols, and preprints, are deposited in public third-party repositories, described in the next section, that assign persistent identifiers. The notebook publishing pipeline operates with a configurable embargo (presently planned at two weeks) between the time an output is finalized in a vault and the time it is rendered to the public site, giving authors a window to add documentation and review before public release.

Data types, metadata, and formats. Research outputs produced at QBI fall into the following categories:

1) source code and job scripts for quantum-chemistry, spin-dynamics, and data-analysis workflows, including Python modules, Jupyter notebooks, and container definitions;
2) input and output files from computational simulations, stored both in code-native formats and in open formats (CSV, HDF5, JSON, plain text) suitable for re-analysis;

3) experimental data from optical and magnetic measurements, stored in instrument-native formats alongside derived open-format exports whenever possible;
4) protocols describing experimental and computational procedures;
5) figures and rendered notebook pages;
6) written materials such as preprints and technical notes.

Metadata is captured at two layers. At the file level, each released output is accompanied by a README and, where applicable, a structured metadata file (for example, a MyST frontmatter block, CITATION.cff, or codemeta.json) that records authorship, contributors, affiliations, funder acknowledgments (using Research Organization Registry identifiers where available), creation and modification dates, upstream dependencies, and a free-text description. QBI provides a README template to project members so that a reader arriving at an output can reconstruct how it was produced and how to reuse it. At the repository level, deposits in Zenodo and other external archives carry Dublin Core-compatible metadata by default, so that outputs are harvestable by external discovery systems.

Publishing venues. QBI publishes outputs in venues chosen to satisfy FAIR principles and to match the established practices of the relevant scientific communities. Reusable code is released on GitHub under the QBI organization. Rendered scientific notebooks are published to the QBI MyST ELN site at https://data.quantumbiology.org. Tagged code releases, datasets, supplementary materials, and any other outputs intended for external archival reuse are deposited to Zenodo, which QBI has selected as its general-purpose external archive: Zenodo is free, CERN-backed, accepts a wide range of artifact types, and assigns a DOI to each deposit, including a concept DOI that always resolves to the most recent version of a versioned record. The Zenodo deposit workflow is being formalized alongside the release pipeline described below. When a discipline-specific data repository would give a particular output greater visibility than Zenodo, QBI uses that repository instead, in consultation with any relevant funder contacts. Protocols are published on protocols.io. Preprints and longer written outputs are posted to bioRxiv, arXiv, or ChemRxiv as appropriate to the subject matter, each of which assigns a DOI at submission.

Release pipeline and automation. QBI is establishing a unified release pipeline that connects the working environment to the external venues described above through its git infrastructure. The intended flow is: a maintainer marks an output for release in its source repository; the pipeline then renders the output, deposits the artifact to Zenodo (capturing the returned DOI), submits the same artifact to DeSci Nodes (capturing the returned dPID), updates the QBI output register, and, after the configured embargo, publishes the rendered output to the public ELN site. The motivation for routing through a single pipeline rather than asking each scientist to deposit to each venue by hand is reliability and ポカヨケ (poka-yoke; mistake-proofing): a manual workflow with three or four steps per output is a workflow where steps get skipped under deadline pressure, and where metadata drifts between venues. Centralizing the release step removes that class of error. While this pipeline is being built out, depositing to Zenodo and DeSci Nodes is performed manually by the Head of Computational Research and Data Infrastructure for outputs that are ready for external release.

Licensing.
Licenses applied. Source code produced at QBI is released under the Apache License, Version 2.0. Research data are released into the public domain under the Creative Commons Zero v1.0 Universal dedication (CC0-1.0). Articles, preprints, technical notes, figures, and protocols are released under the Creative Commons Attribution 4.0 International license (CC BY 4.0). These licenses are applied uniformly across QBI's outputs within their respective categories unless a

specific third-party dependency requires a different downstream license, in which case the more restrictive license of that dependency is honored. These license choices meet or exceed the requirements of open-science funders whose policies mandate permissive, OSI-approved licensing for code and FAIR-compatible dedications for data.

Why Apache 2.0 for code. Apache 2.0 is an OSI-approved permissive open-source license that allows unrestricted use, modification, redistribution, and commercial deployment of code, while preserving attribution. QBI selected Apache 2.0 over the simpler MIT license because Apache 2.0 includes two provisions that MIT lacks: an explicit patent grant from contributors to users, and a retaliation clause that discourages patent litigation over the licensed work. Under that retaliation clause, a user who files a patent infringement suit claiming the Apache-licensed work infringes a patent automatically loses the patent license they themselves received under Apache 2.0, which in practice deters such suits. These provisions give downstream users, including commercial and institutional users, stronger assurance that building on QBI code will not expose them to later patent-based interference. Apache 2.0 is compatible in the downstream direction with almost every other common open-source license (including GPLv3) and is widely accepted by industry, academia, and national laboratories, making it straightforward for collaborators in any of these environments to adopt QBI code.

Why CC0 for data. CC0 is a public-domain dedication rather than a license: by applying it, QBI waives, to the fullest extent permitted by law, all copyright and related rights in the released dataset. Downstream users may use, combine, redistribute, and build upon the data for any purpose, commercial or non-commercial, without any attribution obligation flowing from the legal instrument itself. QBI chose CC0 rather than a Creative Commons attribution license for two connected reasons. First, it removes a well-documented friction point in data reuse known as 'attribution stacking': an analyst who combines many individually attributed datasets can end up with an unmanageable downstream attribution burden, which in practice discourages the kind of large-scale meta-analysis that open data exists to enable. Second, normal scholarly credit for data reuse is handled by the academic citation graph, not by copyright law; CC0 data is routinely cited by convention, and the DOIs assigned to QBI's Zenodo deposits make that citation unambiguous. The combination of CC0 on the legal side and DOI-based scholarly citation on the social side delivers both maximum reuse and appropriate credit to the data producers.

Why CC BY 4.0 for articles, protocols, and other written materials. CC BY 4.0 allows any party to copy, redistribute, remix, and build upon QBI's written outputs for any purpose, including commercially, provided that they credit the original creators, link to the license, and indicate any changes. QBI applies CC BY 4.0 to preprints, technical notes, rendered notebooks, figures, and protocols because, unlike raw data, these outputs embed authorial interpretation and presentation choices for which attribution is both customary and load-bearing: it supports the scholarly citation record, makes reuse auditable, and protects trainees' ability to demonstrate that their contributions have been picked up by the wider community. CC BY 4.0 is the default license for preprints on bioRxiv and for protocols on protocols.io, and it is fully FAIR-compliant.

Intellectual property consequences of the chosen licenses. Under Apache 2.0, QBI retains copyright in the code it produces, but grants to every user a perpetual, worldwide, royalty-free, non-exclusive license to use, reproduce, modify, distribute, and sublicense that code, along with an explicit patent grant covering any patent claims that the code's contributors could license. Downstream users who modify QBI code must note the changes they have made and preserve the license, copyright, and NOTICE files. They are not required to release their modifications under Apache 2.0; Apache 2.0 permits relicensing of derivatives under other licenses (including proprietary ones), subject to the attribution and NOTICE requirements. A user who files a patent infringement suit alleging that the licensed work infringes a patent automatically loses the patent license they themselves received under Apache 2.0.

Under CC0, QBI does not retain enforceable copyright in the released datasets: the dedication waives copyright and neighboring rights to the fullest extent permitted by law, and expressly licenses any residual rights in jurisdictions where a full waiver is not possible, so that downstream users face no copyright-based or database-right obstacle to reuse. This means that QBI cannot, for example, later assert a copyright claim against a party that republishes or commercializes a dataset it has released under CC0. QBI accepts this consequence because it is the point of the dedication: it guarantees to users and to future collaborators that the data will remain unencumbered. Scholarly attribution for reused data is handled by citation of the persistent identifier assigned to each deposit, not by copyright enforcement.

Under CC BY 4.0, QBI retains copyright in its articles, protocols, figures, and other written materials, and grants to every user a perpetual, worldwide, royalty-free license to share and adapt those materials, including commercially, provided that appropriate attribution is given, a link to the license is provided, and any changes are indicated. Licensees cannot apply legal terms or technological measures that restrict others from doing anything the license permits.

Taken together, **these three instruments (Apache 2.0 for code, CC0 for data, and CC BY 4.0 for written outputs) operationalize QBI's commitment to open science**: QBI and its collaborators retain the right to be credited where credit is customary, while imposing no barrier to downstream reuse. Where a funder's contract assigns IP ownership to the funder, these same licenses are the mechanism by which that IP is released to the public to maximize social benefit.

Protection and durability of research outputs. We use a layered strategy to ensure that our outputs are durably identifiable, tamper-evident, and recoverable, even in scenarios where individual services or institutions become unavailable.

Persistent identifiers. Outputs released through external repositories receive persistent identifiers from those repositories at the time of deposit. Zenodo assigns a DOI to each deposit, including a concept DOI that always resolves to the most recent version of a versioned record. Preprints on bioRxiv, arXiv, and ChemRxiv are issued DOIs by those servers at submission. Protocols on protocols.io are issued DOIs. Outputs deposited through DeSci Nodes carry decentralized persistent identifiers (dPIDs) associated with the deposit. The identifier associated with each released output is recorded in the output's README so that downstream citations always resolve to a canonical landing page with complete metadata.

Content-addressed deposits via DeSci Nodes. In addition to repository-based deposits, selected research outputs are published through the DeSci Nodes platform, which associates each deposit with a content-addressed identifier (dPID) backed by decentralized storage (IPFS/Filecoin). This provides two properties that traditional registrars do not. First, the content identifier is a cryptographic hash of the deposit itself, so any later modification to the bytes of an output produces a different identifier; this makes undetected post-publication alteration infeasible and gives readers an independent way to verify that a file they have downloaded is the file that was originally published. Second, the content-addressed layer provides an independent resolution path that does not depend on the continued operation of any single registrar. QBI is extending this workflow so that a broader set of outputs are routinely deposited through DeSci Nodes alongside their repository deposits.

Access control prior to release. Before outputs are ready for public release, working files are held inside the QBI infrastructure described above. The Frida server is not reachable from the public internet: remote access is gated by a WireGuard VPN, and file shares require authenticated accounts. This ensures that pre-release work remains under QBI's direct control until the Principal Investigator determines that an output is ready for public release.

Backup and preservation. All working content on Frida is backed up to Backblaze B2 every six hours. Backblaze B2 is an off-site object-storage provider, and the backups are versioned so that deleted and modified files are retained for a configured retention window. Once outputs are published externally, durability is additionally guaranteed by the preservation policies of the receiving repositories: Zenodo commits to preserving deposited content for the lifetime of the host laboratory (CERN), and the other repositories listed above carry their own preservation commitments. DeSci Nodes deposits are replicated across the Filecoin storage network. In combination, these measures give QBI at least three independent copies of every externally released output (Frida, B2, and at least one external repository), with content-addressed integrity verification for the subset deposited via DeSci Nodes.

Sharing, reuse, and exceptions. QBI's default practice is to release research outputs openly and promptly. QBI does not embargo outputs to accommodate later journal submissions and aligns with its funders' open-access and open-science requirements. Where a funder's policy explicitly requires that funded work not be routed through traditional journal publication (as Astera's policy does), QBI's members comply fully with that requirement for the affected work. Where QBI members collaborate with external groups that intend to submit to journals, QBI's contributions (figures, data, code, and text) are released openly and promptly under the licenses described above, so that collaborators may cite and build on them in their own manuscripts. Exceptions to open, prompt release are limited to situations that involve privacy, security, or biosafety concerns, or license obligations inherited from third-party dependencies. In such cases, the Chief Science Officer, in consultation with any relevant funder contact, determines a handling plan before any affected output is shared. Where a dependency imposes a license incompatible with Apache 2.0, CC0, or CC BY 4.0, the affected output is released under a compatible license, and this choice is documented alongside the output.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. Bonino da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, The FAIR guiding principles for scientific data management and stewardship, Scientific Data **3**, 160018 (2016).

[2] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, Jupyter notebooks—a publishing format for reproducible computational workflows, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (IOS Press, 2016) pp. 87–90.