Omics Data Publishing to International Repositories

V1.2

17 November 2023

Farah Zaib Khan and Jeffrey H. Christiansen

DOI: 10.5281/zenodo.10147703

Contents

Definitions

In this report, the terms listed below are defined as follows:

International / public data repository: refers to the publicly accessible biological data repositories hosted by the European Bioinformatics Institute (EMBL-EBI)[1], the National Center for Biotechnology Information (NCBI)[2], the DNA Data Bank of Japan (DDBJ)[3] and other global 'core data resources' as defined by the Global Biodata Coalition[4] or ELIXIR[5].

Raw data: refers to files coming from an instrument (e.g. a nucleic acid sequencer or mass spectrometer) and generated in a run of nucleic acid sequencing or a mass spectroscopy experiment (e.g. FASTQ read files).

Derived data: refers to processed files (e.g. binary version of a Sequence Alignment/Map (BAM) files, quantitative mass spectrometry data) or derived data artefacts (e.g. normalised read count matrices) that are generated from raw data files.

Data submission: refers to the process of submitting the data (either raw or derived) and accompanying contextual information (metadata) to an international data repository, e.g. registration of studies/samples, data file and metadata preparation and upload, meeting the validation requirements etc.

Data publication: refers to the publication of submitted data and accompanying contextual information (metadata) after validation checks have been passed when the data becomes publicly available.

'Omic data: refers to data that is generated in a high throughput manner using specialised instrumentation. Types of 'omic data include, proteomics, transcriptomics, genomics, metabolomics, lipidomics, epigenomics etc.

[1] ebi.ac.uk/services/data-resources-and-tools (filter list for Data Resources)
[2] ncbi.nlm.nih.gov/guide/all/ (filter list for Databases)
[3] ddbj.nig.ac.jp/index-e.html
[4] globalbiodata.org/scientific-activities/global-core-biodata-resources/
[5] Drysdale et al, The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences, *Bioinformatics*, Volume 36, Issue 8, 15 April 2020, Pages 2636–2642, doi.org/10.1093/bioinformatics/btz959

1. Introduction

1.1 Global context

Scientific data sharing and publication is an integral part of the research data life cycle that is required to disseminate research outputs[6] (see Figure 1). In accordance with the Bermuda and Fort Lauderdale[7] agreements and the more recent Toronto Statement[8], which provide guidelines for scientific data sharing, life science researchers are expected to make their publicly funded 'omic data findable and available for reuse.

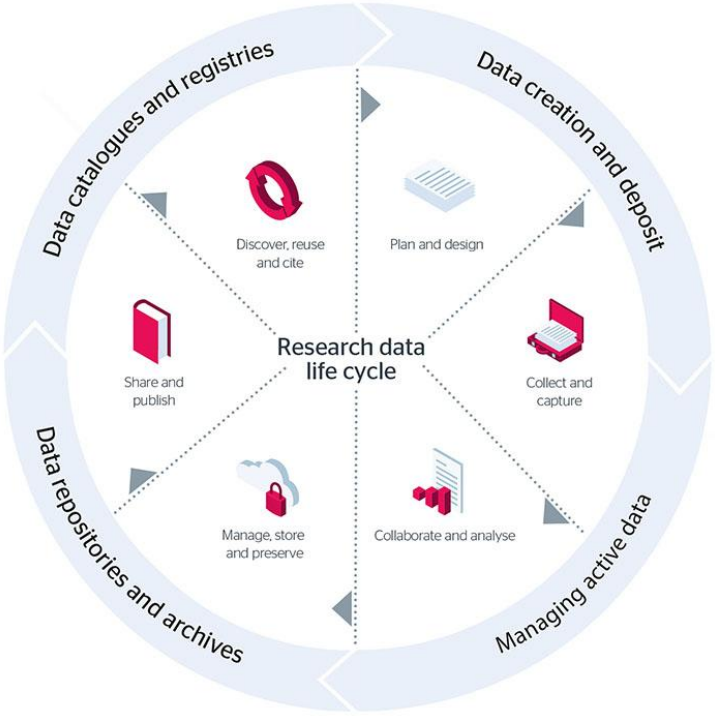

Figure 1. Stages of the Research data life cycle[9]

The standard practice is to submit/publish data to relevant public and freely-accessible repositories, such as the 'core data resources'[10] hosted by the European Bioinformatics Institute (EMBL-EBI)[11], the National Center for Biotechnology Information (NCBI)[12], the DNA Data Bank of Japan (DDBJ) and others[13]. These repositories are primarily funded through their host government or public research monies and provide the global community with long-term access, stable and persistent identifiers for submitted datasets and allow

[6] beta.jisc.ac.uk/guides/research-data-management-toolkit
[7] See Maxson Jones, et al The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *J Hist Biol* 51, 693–805 (2018).doi.org/10.1007/s10739-018-9538-7
[8] nature.com/articles/461168a.epdf
[9] beta.jisc.ac.uk/guides/research-data-management-toolkit
[10] Global Core Biodata Resources globalbiodata.org/scientific-activities/global-core-biodata-resources/
[11] ebi.ac.uk/services/data-resources-and-tools (filter list for Data Resources)
[12] All Resources - Site Guide - NCBI ncbi.nlm.nih.gov/guide/all/ (filter list for Databases)
[13] See globalbiodata.org/scientific-activities/global-core-biodata-resources/ and elixir-europe.org/platforms/data/core-data-resources

public access to the data without paywalls[14]. Making data findable, accessible, interoperable and reusable (FAIR) for fellow researchers following the FAIR principles[15,16] is encouraged or required by an increasing number of funding bodies, as well as being a prerequisite set by many journals prior to publication.

Note that while many public repositories contain what could be considered the 'sole copy' of important biological data globally, a subset of the public repositories also replicate and synchronise some data to provide a redundant global record of these critical research data. The long-standing initiative known as the International Nucleotide Sequence Database Collaboration (INSDC)[17] is an example which operates between the NCBI, EMBL-EBI and DDBJ where information (metadata) about biological studies and samples as well as raw nucleic acid sequence reads, and derived alignment and assembly files are stored across more than one repository (listed in Table 1). This initiative ensures that the data is archived in a redundant manner in geographically dispersed locations (USA, UK, Japan) and that any data hosted at one of the venues is made available via the repositories hosted at the other two venues.

Table 1. Synchronised repositories across DDBJ, EMBL-EBI and NCBI within the INSDC

Data	DDBJ	EMBL-EBI	NCBI
Reads	Sequence Read Archive (SRA)[18]	European Nucleotide Archive (ENA)[19]	Sequence Read Archive (SRA)[20]
Annotated/Assembled sequences	DDBJ Annotated/Assembled Sequences[21]		GenBank[22]
Samples	BioSample[23]		BioSample[24]
Studies	BioProject[25]		BioProject[26]

Most publicly accessible 'omic data repositories welcome the submission of raw or derived secondary data files and associated contextual metadata from the worldwide community and facilitate this by providing:

[14] *Scientific Data's* overview of data repositories: nature.com/sdata/policies/repositories
[15] Wilkinson et al, 2016. The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, nature.com/articles/sdata201618
[16] The FAIR Data Principles - FORCE11 force11.org/info/the-fair-data-principles/
[17] International Nucleotide Sequence Database Collaboration insdc.org/
[18] Sequence Read Archive ddbj.nig.ac.jp/dra/index-e.html
[19] ENA Browser - European Nucleotide Archive ebi.ac.uk/ena/browser/home
[20] Home - SRA - NCBI ncbi.nlm.nih.gov/sra/
[21] DDBJ Annotated/Assembled Sequences ddbj.nig.ac.jp/ddbj/index-e.html
[22] GenBank Overview ncbi.nlm.nih.gov/genbank/
[23] BioSample ddbj.nig.ac.jp/biosample/index-e.html
[24] Home - BioSample - NCBI ncbi.nlm.nih.gov/biosample/
[25] BioProject ddbj.nig.ac.jp/bioproject/index-e.html
[26] Home - BioProject - NCBI ncbi.nlm.nih.gov/bioproject/

(a) submission guides[27,28] (including text and video documentation and tutorials, training and access to support personnel) and accompanying checklists[29] which guide the data submitter (i.e. researcher) to collect and provide at least a minimum amount of contextual information required for sample and data publication, and;

(b) a submission interface (e.g. an interactive user interface, command line or application programming interface (API) submission method) which are appropriate for data submission depending on the scale of the data, frequency of submissions, computing skill set of the individual submitter and the type of data to be submitted[30].

The process of data submission varies depending on the data type, a researcher's preference of repository (if there is a choice), and the mode of submission. Figure 2 depicts a general process that highlights the key steps required for data submission. These steps may differ slightly depending on the repository, nature of the data and submission method. For example, in the case of submitting derived data to the ENA, sample registration is not required unless it is genomic sequence alignment data, in which case the raw reads need to accompany the submission.

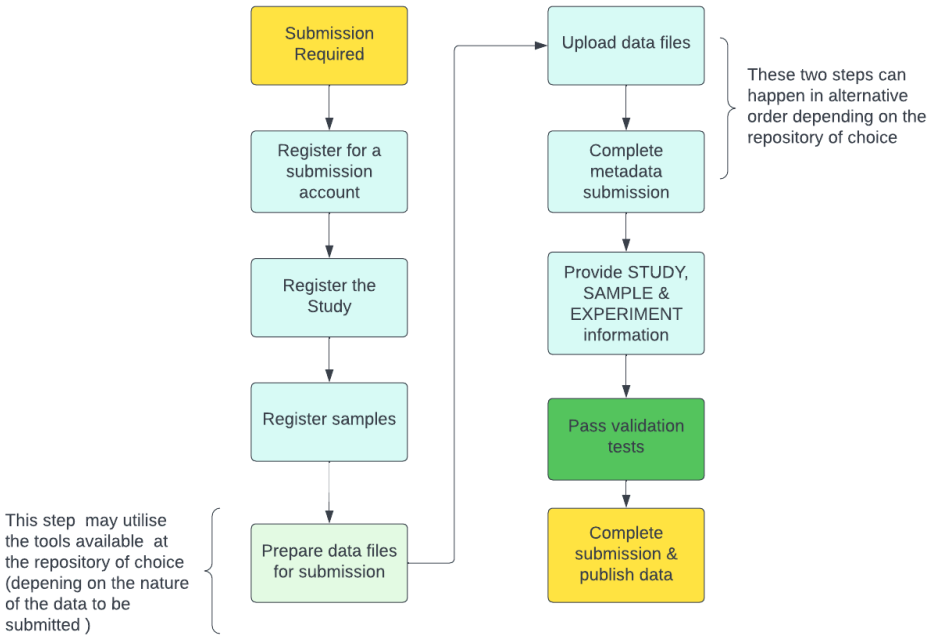

Figure 2. General steps of data submission process. For specific details please refer to the official guides[31,32,33]

[27] Data submission | Services | EMBL's European Bioinformatics Institute ebi.ac.uk/submission/
[28] All Resources - Site Guide - NCBI ncbi.nlm.nih.gov/guide/all/ (filter list for 'Submissions')
[29] For example, see Sample Checklists - ENA Browser ebi.ac.uk/ena/browser/checklists
[30] For example, see Submitting and updating data ebi.ac.uk/ena/browser/submit
[31] Making Submission in SRA Submission Portal ncbi.nlm.nih.gov/sra/docs/submitportal/#6-sra-metadata
[32] SRA File Upload ncbi.nlm.nih.gov/sra/docs/submitfiles/
[33] General Guide On ENA Data Submission — ENA Training Modules 1 documentation ena-docs.readthedocs.io/en/latest/submit/general-guide.html

1.2 The Australian context and aims of this report

During various community engagement activities conducted by the Australian BioCommons engagement team in the past years spanning different scientific communities (including those carrying out genome assembly[34], genome annotation[35], microbiome analysis[36], and comparative genomics[37]), some researchers (n = 20) pointed out that they faced significant challenges during the process of data submission to international repositories. The collective issues includes challenges across:

- identifying the required contextual information or metadata that needs to be collected and then submitted alongside the data;
- formatting the contextual metadata correctly;
- transferring high volumes of data from their local data storage archive(s) to relevant repositories;
- clearing the validation checks placed by these repositories; and
- conducting programmatic submission of data, especially when large numbers of samples and/or high data volumes are involved.

The focus of this report is to:

- further analyse the challenges faced by a broad range of Australian researchers during the data publication phase of the research data life cycle: where data is submitted and published via international data repositories,
- identify the core set of problems these researchers face during this process, and
- recommend potential solutions/strategies the Australian BioCommons can contribute to overcome these problems and streamline data publication to international repositories for Australian researchers.

2. Engagement and analysis methods

The following techniques have been used to understand the past and current state of data submission/publication to various international repositories from Australian researchers.

2.1 Previously Collected User Stories

The Australian BioCommons community engagement team has over the past several years, collected and documented, through a variety of methods (public surveys, focus group sessions, workshops, one-to-one interviews and meetings)[38], the challenges faced by Australian life science researchers using omics data.

[34] Nelson, Tiffanie, & Christiansen, Jeffrey H. (2020). Genome Assembly Infrastructure Roadmap for Australia (4.0). Zenodo. doi.org/10.5281/zenodo.3967970
[35] Nelson, Tiffanie, Griffin, Philippa, & Christiansen, Jeffrey H. (2020). Genome Annotation Infrastructure Roadmap for Australia (4.0). Zenodo. doi.org/10.5281/zenodo.3942716
[36] Nelson, Tiffanie, & Christiansen, Jeffrey H. (2021). Microbiome Analysis Infrastructure Roadmap for Australia (4.0). Zenodo. doi.org/10.5281/zenodo.4978308
[37] Tiffanie M. Nelson, & Jeffrey H. Christiansen. (2022). Comparative Genomics Infrastructure Roadmap for Australia (Version 4.0). Zenodo. doi.org/10.5281/zenodo.7048757
[38] Nelson, TM., Lonie, A., Gustaffsson, J. and Christiansen, J. The Australian BioCommons Community Engagement Strategy: Engaging Researchers at a National Scale to Understand Challenges and Deliver Solutions, eResearch, 2020. doi.org/10.5281/zenodo.4158499

Data collected through surveys included specific questions inquiring about a researcher's experience in making their data publicly available, such as *Do you make your datasets publicly available? Where do you make the datasets available? Have you encountered any difficulties in making datasets available? If you don't make your datasets publicly available, why not?*

The challenges faced were documented as individual "user stories"[39], a method of data and information collection and storage that is part of the Agile Framework[40] of work planning and execution. User stories allow complex needs to be captured in a short-form that describe the challenge or desired capability written in simple terms and have a repetitive format structure. The requirement is summarised in the format: As a [*User type*], I require [*a function or tool*], so I can [*achieve a goal or overcome a current issue*]. For example, *As a researcher, when I am about to publish a paper I require clearer instructions as to what is required to submit data to a relevant data repository, so I can get my data submitted and my paper published*.

For this report, we have used the many hundreds of user stories collected thus far by the Australian BioCommons community engagement team as a starting point to identify the issues and roadblocks with the data submission/publishing process. We filtered all the user stories collected to identify those specifically mentioning issues with the submission process to the international repositories[41] either hosted by NCBI, DDBJ or EMBL-EBI (n = 20). User stories indicating data publication elsewhere e.g. organisation servers or generalist cloud storage solutions (e.g. Zenodo[42] or figshare[43]), or solutions offered by journal publishers (e.g. GigaDB[44]) etc. (n = 9) are out of scope for this analysis (See Appendix 1). These user stories along with a review of both the literature and the requirements described for submission to data repositories hosted by NCBI, EMBL-EBI or DDBJ guided the questionnaire design described in the next section.

2.2 Interviews

The focus of the previous community engagement activities (public surveys, focus group sessions, workshops, one-to-one interviews and meetings) was broad and not designed to explore the specific aspects of data publication in detail. As a result, the contextual details available in previously collected user stories were not adequate enough to understand the entire problem space.

We therefore decided to conduct in-depth one-on-one interviews with selected researchers who had identified challenges in data publication to NCBI, EMBL-EBI or DDBJ hosted repositories in previous engagement activities. A script (See Appendix 2) was developed to guide these in depth interviews which explored the scale of experimental data (to be submitted), local metadata and contextual information, data management methods and systems in use by these researchers, as well as the challenges that the

[39] A user story is the smallest unit of work in the Agile Framework of project planning and work execution atlassian.com/agile/project-management/user-stories
[40] Agile Framework is a method applied to project management and product planning, orginiall designed for software creation and improvement atlassian.com/agile
[41] Relevant repositories
- EBI: BioSamples, BioStudies, ENA, MetaboLights, PRIDE, UniProt
- NCBI: GenBank, Sequence Read Archive (SRA), GEO, BioProject, BioSample
- DDBJ: Annotated/Assembled Sequences (DDBJ), DDBJ Sequence Read Archive (DRA), BioProject, BioSample MetaboBank, Genomic Expression Archive (GEA)
[42] Zenodo: zenodo.org/
[43] figshare: figshare.com/
[44] GigaDB: gigadb.org/

researchers faced during the submission process. The design of the script and questions was informed by the insights gleaned from the initial set of user stories outlined in Section 2.1 as well as related literature including the documentation and usage guides available on data hosting repositories. Four researchers accepted our invitation to participate in these interviews.

2.3 Data Chaperoning Requests

To supplement the information gleaned through the small interview sample size described in Section 2.2 (n=4), we elected to also interview staff who operated the EMBL-ABR Data Chaperoning Service[45] to further understand the challenges faced by the community .

The EMBL-ABR Data Chaperoning Service was active between 2016 and 2019, and was operated by QCIF[46] on behalf of the EMBL-Australia Bioinformatics Resource (EMBL-ABR - a precursor of the Australian BioCommons)[47] to support researchers who required assistance with the process of data preparation and submission to various international repositories hosted by NCBI or EMBL-EBI. During its lifetime, the EMBL-ABR Data Chaperoning service helped more than 50 researchers across Australia to submit 2,012 files relating to 1,119 samples to databases hosted by both EMBL-EBI (i.e. BioStudies, BioSamples, ENA, MGnify and UniProt) and NCBI (i.e. SRA, GEO, GenBank). A similar service has continued informally for Queensland based life science researchers since 2019 and is operated by QCIF Bioinformatics[48].

Due to its nature of providing help to a large number of researchers throughout the entire data publication process to a variety of EBI and NCBI hosted databases, the operators of this service are well placed to understand bottlenecks faced by the community. The requests contained key information about the hurdles typically encountered by the researchers attempting to submit a variety of data (e.g. RNAseq, nanopore sequencing reads, PacBio raw sequencing data, genome assemblies etc.) as part of their publication process. We therefore also interviewed the staff responsible for operating the Data Chaperoning Service as a proxy to understand the challenges faced by the many clients of the Service.

2.4 Bioplatforms Australia Data Portal (BPA-DP)

The BioPlatforms Australia (BPA) framework initiatives[49] are national projects utilising integrated omics infrastructure to generate omics-based reference datasets and knowledge supporting researchers in various fields including agriculture, biomedicine and environmental science. The resulting datasets along with the associated structured metadata are stored in the Bioplatforms Australia Data Portal (BPA-DP) which has been built to support pre-publication data sharing within each Framework project[50]. At an appropriate time in each project, data is submitted to various international repositories, primarily NCBI-SRA, and more recently MetaboLights[51].

---

[45] web.archive.org/web/20190707225554/https://www.embl-abr.org.au/data-chaperoning/
[46] qcif.edu.au/
[47] See Schneider et al, Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia, Briefings in Bioinformatics, Volume 20, Issue 2, March 2019, Pages 384–389, doi.org/10.1093/bib/bbx071
[48] Bioinformatics & Biostatistics – QCIF qcif.edu.au/services/bioinformatics-and-biostatistics/
[49] bioplatforms.com
[50] data.bioplatforms.com
[51] ebi.ac.uk/metabolights/

We interviewed both the BPA-DP Operations staff (who manage submissions of raw nucleic acid sequencing data from the BPA-DP to NCBI-SRA on behalf of various framework projects) as well as an individual who has published 100's of metabolomics datasets from the BPA-DP to MetaboLights on behalf of the Sepsis framework initiative[52] and raw nucleic acid sequencing data[53] to NCBI-SRA on behalf of the Genomics for Australian Plants project[54] to further understand challenges these individuals have faced when submitting data to international repositories.

3. Findings & Discussion

3.1 General Notes

– Omics DataType

Most interviewees were focussed on submission of nucleotide sequencing data with the exception of the data chaperoning service which also handled a small number of submissions to UniProt and the data submissions to MetaboLights from the BPA-DP.

– Repository Preference

- In the case of data chaperoning requests, most researchers did not have any preference for a particular repository within the INSDC and sought advice from the operators of data chaperoning services about the choice of repository (e.g. at EBI or NCBI).

- Despite challenges faced during their data submission experience, one of the interviewees preferred an NCBI hosted repository (GenBank), but were also open to trying EBI hosted repositories in future. The reason for choosing a particular repository was mostly based on the existing protocols/methods in place in a particular lab/institute or a researchers' prior experience with the submission process of a repository. In some cases, interviewees suggested they chose a particular repository because the access was easier, more efficient or capable for the data access.

- In the case of data chaperoning service requests, the researchers generally relied on the advice provided by the operators of the service regarding the choice of an appropriate repository for their data. The operators of the data chaperoning service personally preferred submission to EBI hosted repositories because they found that the documentation was clearer, and the process simpler, when compared to NCBI.

[52] Sepsis - Bioplatforms bioplatforms.com/projects/sepsis/
[53] ncbi.nlm.nih.gov/bioproject/PRJEB49212
[54] Genomics for Australian Plants genomicsforaustralianplants.com/

– Data Categories

The data to be submitted can be divided into two categories: raw data (e.g. FASTQ[55] read files) and derived data (e.g. processed sequence files as BAM[56] alignment files or derived data artefacts as normalised read count matrices).

3.2 Common Themes

T1 – Data Submission is Often Unplanned in Experimental Procedures

For some researchers, the generation of data, such as nucleotide sequence or other omics data, and its ongoing access, interpretation and findability is foremost in their mind. This is partially because the time and effort, including the dollar value, to generate the data and bioinformatically convert it into meaningful products is costly. Some researchers (especially early career researchers) may not be aware that data submission/publication to an international repository is a requirement upon the submission to a peer-reviewed journal and increasingly a requirement of many funding bodies and academic institutions. It is well-understood that considering the data and metadata requirements for data submission at the beginning of a research project is more efficient and better than tackling these challenges *post-hoc*.

Meeting minimum metadata requirements and complying with the specific metadata formats required by various data repositories is a critical factor in completion of the data publication process. When researchers consider data submission as an afterthought, instead of adopting a methodical approach to capture and represent the appropriate and required contextual information throughout the project, it can result in incomplete metadata capture and make it challenging to complete the submission of a high quality data record.

The BPA-DP hosts data generated from many framework initiative projects. The project managers of these framework initiatives ensure the collection of all the required study, sample and library metadata that will be required for submission to international repositories from the beginning of the dataset's life cycle. The project manager supports and encourages the capture of metadata by providing the knowledge to scientists and researchers involved in the project about metadata capture and requirements from a repository, funding and publishing perspective. This includes a strict file naming convention which is to be followed by the researchers to support the automatic submission from the BPA-DP to the relevant NCBI repository (SRA)[57]. The managers also provide comprehensive templates in MS-Excel for metadata recording that must be completed fully prior to any omics data generation. The templates include the minimum metadata that is required for submission to appropriate international repositories. The fields, 'Sample' and 'Study', are imported directly into the BPA-DP from the Excel templates and then associated with related omics data files upon their upload to the BPA-DP. The enforcement and adherence to metadata guidelines and strict naming conventions, make the submission process from the BPA-DP to NCBI easier, highlighting that planning for data publication at the beginning of the research data life cycle

[55] FASTQ file: refers to a text file that contains a genomics sequence read data along with sequence read identifier and other information about the quality of the sequence read, such as per-base quality scores, ncbi.nlm.nih.gov/sra/docs/submitformats/#fastq-files

[56] BAM or Binary Alignment/Map file: refers to file that is a compressed version of the Sequence Alignment/Map (SAM) format, ncbi.nlm.nih.gov/sra/docs/submitformats/#bam-files

[57] github.com/BioplatformsAustralia/bpa-submission-generator

can mitigate the challenges associated with incomplete metadata capture.

T2 – Significance of Contextual Information

Amongst some researchers, there is a lack of understanding about the significance of the contextual information/metadata to be made available for data publication and its downstream impact on data reusability, reproducibility and findability. The lack of motivation amongst researchers to spend resources (time, especially) to record and provide detailed contextual information was observed by operators of the data chaperoning service and one of our interviewees who has several years of experience in both bioinformatics analysis and data publication interactions with several international repositories. The lack of awareness about the impact of accompanying metadata on data re-use, discoverability and overall reproducibility of an experiment[58] results in less interest on part of researchers in putting efforts and resources to capture and record contextual information[59].

T3 – Metadata Requirements

In some instances, the researchers had limited knowledge of the contextual information/metadata requirements of a repository before carrying out the analysis/research. This resulted in incomplete metadata capture and subsequent failure to pass the validation checks in place by the dedicated repositories.

T4 – Lack of Appropriate Local Data Management Systems

Contextual metadata is collected at different stages of the data lifecycle: this means that more than one team/individual/resource (e.g. research labs, sequencing facilities, existing databases etc.) are involved in handling the collection of various metadata elements. The submission and publication of the data requires metadata collation from all these teams/individuals/resources. Unfortunately, most researchers do not have access to an adequate unified item-level data management system to help structure their data files and accompanying metadata, throughout the research lifecycle. One of our interviewees identified the importance of having a well-defined, open-source system to handle storage and documentation of metadata throughout the data lifecycle, but also noted that they did not have access to such a system themselves. Examples of existing data management systems that are used by various organisations globally and which could be explored in future for omics data management are MediaFlux[60], iRODS[61], GeneStack[62], CKAN[63], Gen3[64] and CyVerse[65].

T5 – Correct Metadata Template

In some cases (particularly in data chaperoning requests), life science researchers requested that repository/platform operators provide suitable metadata templates:

---

[58] ardc.edu.au/news/enabling-and-enhancing-the-discovery-and-reuse-of-data-with-metadata/
[59] Rajesh, A., Chang, Y., Abedalthagafi, M.S. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol* 22, 106 (2021). doi.org/10.1186/s13059-021-02332-z
[60] arcitecta.com/mediaflux/features/
[61] irods.org/
[62] genestack.com/products/omics-data-manager/
[63] ckan.org/
[64] gen3.org/
[65] cyverse.org/

> *"I'm not exactly sure how to format this. If it is possible, would you please be able to supply an example of what the metadata should look like?"*

The data chaperoning team provided the researchers with Excel sheet checklists that contained both the mandatory and optional metadata fields. Note that these checklists[66] are available from the relevant repositories.

T6 – Correct Repository

Some life science researchers also required help from the data chaperoning team when choosing the right repository for the data they aimed to submit:

> "*If you might be able to give any advice regarding submission to either ENA or GEO? This is my first time having to submit data into a database, so I'm just not really sure where to start*"

T7 – Registering a Study or Sample

Researchers generally found the process of registering studies and samples straightforward, provided that they also had the required metadata in hand. For EBI hosted repositories (e.g. ENA, MetaboLights etc.), the registration process for studies[67] and samples[68] can be carried out interactively as well as programmatically, which caters to users with varied IT skills. At NCBI, an online browser-based wizard[69], with a built-in taxonomy browser, provides a step-by-step guide for sample and project/study registration. The user experience varied depending on the repository of their choice and is detailed later in this report.

T8 – Preparing Files for Submission

The results of interviews as well as earlier user stories did not generally indicate any issues faced when preparing raw and derived data files for submission to either EBI or NCBI hosted repositories. One exception is detailed below in the "Derived Data" section where researchers found it challenging to produce the .sqn files using dedicated software (available from NCBI) for submission to GenBank.

T9 – Raw Data

Based on our analysis, submission of raw sequencing data files (termed as "raw reads" in ENA metadata model[70] and "Experiment" and "Run" in the NCBI/SRA data model[71,72]) is more straightforward than derived data publication. This appears to be because raw data can be submitted interactively through web-based portals[73] and the validation requirements for raw data are also fewer than derived data.

---

[66] For example, see Sample Checklists - ENA Browser ebi.ac.uk/ena/browser/checklists
[67] ena-docs.readthedocs.io/en/latest/submit/study.html
[68] ena-docs.readthedocs.io/en/latest/submit/samples.html
[69] ncbi.nlm.nih.gov/sra/docs/submitbio/
[70] ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html
[71] Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. doi: 10.1093/nar/gkq1019. Epub 2010 Nov 9. PMID: 21062823; PMCID: PMC3013647.
[72] SRA Metadata and Submission Overview ncbi.nlm.nih.gov/sra/docs/submitmeta/#anatomy-of-the-sra-data
[73] https://ena-docs.readthedocs.io/en/latest/submit/reads.html

T10 – Derived (Secondary) Data

Strict validation checks were encountered by researchers who attempted to submit derived data (e.g. genome assemblies, targeted sequences, read alignments and sequence annotations, termed as "analysis" in ENA metadata model and NCBI-SRA data model) to data repositories. Error messages received when these automated validation checks failed can be cryptic, and interviewees have found that the documentation required to interpret these messages may be either missing or inadequate to identify the problem. This leaves the researchers with little guidance on how to troubleshoot or proceed further.

One of the interviewees contacted the support team at NCBI to clarify the meaning of error messages they encountered when submitting genome annotations (sequence feature). This researcher was using the software tbl2asn[74] and encountered several errors, even though the input file guidelines provided by NCBI were followed. In this case, the wait time for a response from the support team was more than two months, and by this time the researcher had independently resolved the error and progressed further, only to face new error messages. After attempting to resolve the issues on their own, the researcher ended up submitting the unannotated genome assemblies to NCBI and made the analysed/result data available in a generalist data repository i.e. Zenodo.

An important factor mentioned by several interviewees specifically for annotated genomes is the wait time for getting the annotations approved by the repositories. Submitting the annotated genomes to NCBI Genbank typically took one month and in some cases, eight months with a trail of emails between the researchers and the team at NCBI/Genbank.

Another complicating factor for derived data submission is that the submission of annotation data is only possible programmatically[75] (i.e. via the command line and not via the interactive Graphical User Interface (GUI) method in the web browser). This makes it challenging for users who have limited expertise with the required programmatic methods.

T11 – User Experience

Interviewees stated that the GUI available for data submissions to some repositories was not intuitive, which results in a compromised user experience. One of the interviewees, who has been submitting data to a number of NCBI hosted repositories for several years, acknowledged that the data submission GUI has improved over the years and the process is now easier. The data chaperoning service operators confirmed that a strategic effort aimed at improving user experience across EMBL-EBI databases[76] has resulted in GUI improvements for many EMBL-EBI hosted repositories, and they are now more intuitive to use.

An attendee at the Australian BioCommons metabolomics community meetings[77] expressed sufficient challenges with submitting data to MetaboLights which resulted in them electing not to submit any data at the end of their project. They indicated that the current interface does not allow for modifications or edits, making it impossible to correct errors on the spot. They ultimately selected an alternative repository for publishing data (Metabolomics Workbench[78]). Another interviewee shared these sentiments and stated

[74] ncbi.nlm.nih.gov/genbank/tbl2asn2/
[75] ena-docs.readthedocs.io/en/latest/submit/analyses.html
[76] ebi.ac.uk/about/teams/web-development/
[77] biocommons.org.au/events/metabolomics-community
[78] https://www.metabolomicsworkbench.org/

that their data still includes an error code after validation and that the documentation does not describe how to resolve the error.

T12 – Documentation

Most researchers we interviewed felt that the documentation provided by many repositories could be more user friendly. The information is perceived to be presented in convoluted formats and the user is expected to dedicate a significant amount of time to extract the information that is pertinent to their particular task. This was a significant concern highlighted in the initial user stories and was reflected in nearly all the interviews conducted for this report.

4. Recommendations

The issues revealed by the interviews, user stories and other interactions point to a need to better support Australian life science researchers who wish to make their data public through submission to various international repositories. We believe that while some improvements can be made at a local level, the greatest impact will result from strategies implemented at a global level and that efforts to improve the data submission and publication process should be carried out in close collaboration with the organisations that host the international repositories.

To highlight the scope of several proposed solutions outlined below, we have described the anticipated users of these solutions with four different personas[79]. Personas are a powerful usability research tool that are utilised to understand the motivations and needs of our target users, and the solutions that may be needed to address their requirements.

Our personas are:

1. Leia, The Life Scientist,
2. Beena, The Bioinformatician,
3. Cameron, The Computer Scientist, and
4. Brad, The Beginner Biologist.

These four user personas are characterised by varied skill sets and require tailored support for the different aspects of the data publishing process (see Appendix 3). Our interviewees, and the researchers who submitted data chaperoning requests, can all be classified using one of these personas. Our recommendations in this section aim to address the needs of these user personas. However, each solution is not necessarily suitable for all personas.

The proposed solution space is divided into three stages depending on the time and planning resources required to implement each solution.

- Short-term: solutions that could be implemented within a year;
- Medium-terms : solutions that could be implemented in 1-2 years, and;
- Long-term: strategies and solutions categorised as overarching goals that could be achieved in a longer period of time (>2 years)

Table 2 contains our proposed solutions to the community challenges detailed in this report. The proposed solutions are all currently (as of Q4 2023) directed towards the repositories hosted by EMBL-EBI. This is because the Australian BioCommons and ELIXIR[80] have an existing collaboration agreement[81] in place, which enables cooperation to address challenges of international scope around several areas including data, training, tools, and interoperability standards. EMBL-EBI is a node of ELIXIR, and via the collaboration agreement, connections have already been established with some of the EMBL-EBI repository management teams (e.g. MetaboLights and ENA). Over time we expect the Australian BioCommons to develop and strengthen relationships with NCBI and/or DDBJ, which will then allow us to propose further NCBI- or DDBJ-specific solutions.

[79] interaction-design.org/literature/article/personas-why-and-how-you-should-use-them
[80] ELIXIR Europe elixir-europe.org/
[81] ELIXIR - Australian BioCommons Collaboration Strategy biocommons.org.au/elixir-collaboration

Table 2. Recommendations, their reasoning, the concerned teams to be involved in materialisation of the recommendations and the relevant user personas

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
Short	Brad The Beginner Biologist Leia The life scientist	**R1- Deploy ENA upload tool within Galaxy Australia:** Install, test and promote the ENA-upload tool within Galaxy Australia as a service to allow researchers to submit raw nucleotide sequence reads or genome assembly files to the European Nucleotide Archive (ENA) through their web browser.	This Galaxy-based tool was originally developed by ELIXIR-Belgium to enable easier submission of COVID-19 viral sequence data to ENA by researchers with little to no computational background. Galaxy Australia is the established flagship service of the Australian BioCommons that is used by 20,000+ researchers for a variety of bioinformatics tasks. It is a professionally managed service that can act as the natural home in Australia for an ENA upload capability (and potentially other data submission tools into the future). Since submission of nucleotide sequences (raw reads) is most required, supporting a tool facilitating this submission type will help to address the needs of a large audience.	**Galaxy Australia team:** To install the ENA Upload Tool, check the validity of the tool on Galaxy Australia (e.g. to ensure that user credentials required for submission of data to ENA peacefully co-exist with any other credentials required by Galaxy Australia), and to ensure that the system is tested thoroughly. **BioCommons Community Engagement (Business Analyst, BA):** To test the working of the ENA upload tool with example datasets (noting that this tool supports submission of test datasets to the ENA Test server[82]).
Short	Brad The Beginner Biologist Leia The life scientist	**R2 - ENA upload tool support documentation:** Tailor the help documentation of the ENA-upload tool to guide researchers specifically aiming to use the Galaxy	Documentation exists elsewhere (e.g. in ELIXIR Belgium's RDM guide and the global Galaxy training network) however these were originally documented for Galaxy Europe. These documents should be tested on Galaxy Australia and	**BioCommons Community Engagement (BA):** To lead the deployment of relevant documentation alongside Galaxy Australia

[82] wwwdev.ebi.ac.uk/ena/

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		Australia hosted instance of the ENA-upload tool. The existing documentation (e.g. RDM guide and Galaxy training) should be utilised as a starting point to create a generalised version of the documentation that is relevant to Galaxy Australia and is available either via the Galaxy Training Network or elsewhere.	modified if necessary to align these with Galaxy Materials in the Galaxy Training Network function both as user documentation that can be pointed to from within a Galaxy instance, as well as being material for either training workshops or self-paced training.	team. **Galaxy Australia team:** To assist with deploying relevant documentation through the most appropriate means within the Galaxy framework.
Short	Brad The Beginner Biologist Leia The life scientist	**R3 - Workshop on how to use the ENA upload tool in Galaxy Australia** There are a few crucial steps in the process of submitting data via ENA upload tool such as adding ENA Webin credentials to a user's Galaxy account, inputting metadata interactively or via a metadata template and a final submission of the reads to ENA. This workshop will help researchers follow an example with the Galaxy Australia team to make a submission to the test ENA server.	Having documentation tailored for Galaxy Australia (suggested in R2) followed by the workshop will equip the researchers with the necessary information to utilise the ENA upload tool for their submissions. R1, R2 and R3 together will improve user experience (identified in T11) for researchers aiming to submit raw data by providing an alternative interface that we can customise and improve iteratively by incorporating feedback from users.	Potential trainers: • Galaxy Australia team, • BioCommons Community Engagement (BA). BioCommons team to coordinate the event: • Training • Communications • Community Engagement
Short Medium	Brad The Beginner Biologist Leia The life scientist Beena The Bioinformatician Cameron The computer scientist	**R4 - Series of targeted webinars explaining how to structure and record biological sample and experimental metadata and submit data to various international repositories:** Webinars (typically 40-45 minutes) are envisaged for awareness raising, which will address either general topics (e.g. data	We interviewed some individuals and teams who have considerable experience in making data submissions on behalf of other researchers. Inviting these individuals as well as EBI team members who manage various public repositories to deliver webinars will make it possible to deliver key information that may not be	Potential speakers: • Domain experts • Data chaperoning team at QCIF • EBI team members • Frequent submitters BioCommons teams to coordinate the

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		management steps[83]), or repository-specific, domain-specific, data type-specific, or submission mode-specific topics. For example, guiding researchers to: choose the appropriate repository for their data, select appropriate minimum information specifications for metadata description, describe the requirements of a specific repository, share personal experiences for peculiar hurdles (errors, requirements) when dealing with a specific data type (e.g. genome annotation, metabolomic data etc.). The webinars can be recorded and made available for future reference. These webinars can be organised under the auspices of the existing Australian BioCommons[84], and potentially EMBL-EBI [85] training programs.	apparent elsewhere. Researchers from anywhere in Australia will be able to attend virtually making it inclusive. The recordings of these webinars will be findable and viewable via YouTube and will be a valuable artefact for future reference.	event: ● Training ● Communications ● Community Engagement
Medium	Brad The Beginner Biologist Leia The life scientist	**R5 - Series of workshops** demonstrating how to submit data to various international	Since the format of webinars only allows one topic of interest to be covered in a concise fashion and with limited	Potential trainers: ● EBI team members

[83] Data Management in Simple Steps | RDM Guide rdm.elixir-belgium.org/data_management_steps
[84] Find a webinar or workshop — Australian BioCommons biocommons.org.au/training
[85] ebi.ac.uk/training

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
	Beena The Bioinformatician Cameron The computer scientist	repositories[86] The webinars in R4 can optionally lead to a series of workshops (depending on the feedback and demand from community), each dedicated to cover extended and hands-on aspects of the topic of interest. The duration of these workshops can vary from 2-3 hours (or longer if required) where the audience is actively involved in hands-on practical activities during the session. These workshops should be conducted online to ensure equitable access to researchers from anywhere in Australia.	interaction from the audience, researchers and participants will benefit from the opportunity to follow up on topics raised in the webinar. These follow up queries along with other related topics can be well-covered in a workshop format as it will involve active involvement from the audience with the opportunity for interactive Q/A sessions. The participants will have the opportunity to undertake practical exercises and interact openly with the instructors/facilitators for active learning. The series of workshops can build on one another, making it possible to cover a topic in depth if required. It is noted that not all webinars outlined in R4 would necessarily have an associated workshop to follow on and we expect that this will depend on the demand from the audience. Together R4 and R5 will directly target the issues identified in common themes e.g. T1, T3, T5-T8.	• Domain experts • Data chaperoning team at QCIF • Frequent submitters BioCommons team: • Training • Communications • Community Engagement
Medium	Leia The life scientist Beena The Bioinformatician	**R6 - Guides which help to explain how to structure and record biological sample and experimental metadata and**	Having a concise document to refer to when a researcher begins a research project is going to be beneficial as it will	Community Engagement team (BA):

[86] Note: In June 2021 the Australian BioCommons started a conversation with members of the ENA team (specifically Sam Holt, the ENA training manager) regarding organising training events re. Data submission to ENA specifically for Australian researchers. Dr Holt was supportive of the idea but unfortunately the timings of the conversations coincided with his departure from ENA, hence those training efforts could not materialise. Australian BioCommons is eager to re-establish the working relationship again with the current training team at EBI.

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		submit data to various international repositories: Concentrated, yet high-level guides summarising lengthy documentations (e.g. the ENA training modules) to provide targeted information about a topic of interest. These guides could follow the format of "Ten simple rules .. " with steps / bullet points/ rules that should be followed to achieve the desired outcome. These guides should include the flowcharts for the end-to-end process of a specific activity. For example, if one needs to submit their data to MetaboLights, the appropriate guide should provide the researcher with the steps to be followed to successfully submit their data. These guides should be produced in close collaboration with the repository management teams. We recommend these guides to be findable via each respective repository as these artefacts would be of interest to researchers globally.	help researchers to think about data submission throughout the data life cycle without much of a learning curve or without going into detailed documentation presented by the repositories. The Ten simple rules format is engaging and succinct[87] and will convey key information especially when written in collaboration with the EBI repository management teams. Including graphical representation e.g. flowcharts will also further aid the effective communication of key information related to the topic of interest. We have had preliminary conversations with the MetaboLights team[88] who are open to receiving feedback on the existing documentation required for data submission to MetaboLights that we received while interacting with the community.	• To lead documentation design Partners: • EBI repository managers
Medium	Brad The Beginner Biologist Leia The life scientist Beena The	**R7 - Documentation improvement:** Collaborative improvement where necessary of the documentation of various	Feedback we have received from a variety of researchers and other stakeholders documented in this report strongly suggest the improvement or streamlining of the	Community Engagement team (BA): • To lead documentation design • Response to roadmaps DevOps

[87] Dashnow, H., Lonsdale, A., & Bourne, P. E. (2014). Ten simple rules for writing a PLOS ten simple rules article. *PLoS computational biology*, *10*(10), e1003858. doi.org/10.1371/journal.pcbi.1003858

[88] twitter.com/AusBiocommons/status/1550327273349267456

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
	Bioinformatician Cameron The computer scientist	international repositories (e.g. MetaboLights, ENA, etc) based on the feedback gathered from the community regarding the submission process, metadata requirements, metadata format, transparent validation criteria and appropriate submission mode for different types/scale of data. The ideal scenario would be to work in collaboration with the repository management teams and aim to incorporate improvements directly into the respective repository's official documentation instead of replicating and hosting this information elsewhere. This will make the information accessible and visible to anyone aiming to publish their data via EBI.	existing documentation hosted by multiple repositories is required. Since each repository should be the first/primary source of information to be consulted by the researchers during the submission process, it is imperative that the feedback from the scientific community be incorporated to make the documentation user friendly, accessible and clearer. If Australian BioCommons elects to host additional documentation guides information locally e.g. the BioCommons website or other local forms of documentation, this information would unlikely be found or accessed either globally or widely by Australian researchers. Overall implementation of R6 and R7 will resolve issues identified in several themes e.g. T3, T5, T6, T7, T8, T10 and T12.	role Partners: • EBI repository managers
Long	Brad The Beginner Biologist Leia The life scientist Beena The Bioinformatician Cameron The computer scientist	**R8 - Contribute feedback about submission interface improvements from Australian Users to the respective International repository teams:** Initiate discussions with the repository management teams (e.g. MetaboLights, ENA, etc) to contribute user suggestions for improving the submission interfaces for	This activity would require in depth research of the interfaces of various repositories with use cases tested with the researchers to identify the exact points of pain and issues to be targeted for improvement. Hence it is termed as a long term goal that should be focused on after raising awareness and training researchers of the requirements and	Community Engagement team: • Initiate discussions • Suggest the changes (extracted from the findings when interacting with researchers and conduct further research to identify the specific issues) • Arrange community-led testing

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		improved user experience. The feedback from various avenues (meetings, interviews, user stories) echoed similar concerns about various submission interfaces not being intuitive. The repositories would benefit from improvement of the user experience to make the process of submission easier. We have encountered several cases where researchers gave up on publishing their datasets due to the compromised user experience. Gathering further feedback from these researchers and relaying that information to the repository management team could help guide these changes to the submission interface.	principles of data management in general through our earlier recommendations (R4-R7). In addition, getting EBI repository teams on board is crucial to achieve this goal. The Australian BioCommons through our community engagement approach can identify and harness the collective effort of many different user types from across Australia to help inform this work. As stated earlier, the MetaboLights team is already receptive to the feedback from our communities who engaged in the data submission process. Changes in submission interfaces will address the concerns echoed in T10 and improve the user experience detailed in T11.	when/if the submission interface is modified. Partners: • EBI repository managers
Short Medium	Brad The Beginner Biologist Leia The life scientist Beena The Bioinformatician Cameron The computer scientist	**R9 - Improving findability of pertinent documentation from the Australian BioCommons website:** As described in R6 and R7, we aim to work closely in collaboration with international partners to produce/improve the documentation reflecting the feedback from our scientific communities. However, we recommend revising our own website (https://www.biocommons.org.au/) and local documentation to include direct	Accumulating information about the existing resources (e.g. help documents, webinars and training content) and presenting it interactively through BioCommons website (and from the most relevant page(s) of the website) will improve the accessibility and findability of such resources. It will help researchers access multiple resources through various points of entry.	Community Engagement team: • To lead documentation design Comms team: • Make changes to website

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		references/pointers to existing resources (e.g. recorded webinars[89], detailed documentations[90] etc.) and newly designed artefacts (as a result of implementing R6 and R7).		
Long	Leia The life scientist Beena The Bioinformatician	**R10 - 'BYOD' sessions:** Some researchers in our consultation, especially those aiming to submit genome annotation data, expressed unique challenges (e.g. strict validation checks and no venues to submit the manually curated genomes). We suggest organising semi-regular 'Bring Your Own Data (BYOD)' sessions for researchers requiring to submit genome annotation data where relevant staff from repository management teams can guide these researchers in tackling uncommon/distinctive errors. We recommend ensuring pre-registration for these events and the data should be shared with the repository team in advance.	During our engagement activities, it became evident that the challenges for derived data submission/publication where extensive manual and automated validation checks were in place, are significant, and to help users would require more attention than a webinar or a conventional workshop which uses predefined data files. Both formats can help to address the fundamental concepts related to the genome annotation data submission but for peculiar issues/errors, specific BYOD sessions could help the researchers understand these errors and resolve the issues with the help of concerned repository team members. It is worth noting that this activity will only be possible with the agreement and ability to resource these activities from the various EBI team members. These sessions will particularly aim to target the issues identified in T10.	Potential speakers: • EBI team members • Data Chaperoning team at QCIF BioCommons team: • Communications & Training • Community Engagement

[89] https://www.ebi.ac.uk/training/events/metabolights-home-metabolomics-experiments-and-derived-information/
[90] https://ena-docs.readthedocs.io/_/downloads/en/latest/pdf/

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
Long	Brad The Beginner Biologist Leia The life scientist Beena The Bioinformatician Cameron The computer scientist	**R11 - Investigating a role for the Australian BioCommons as a Data Broker:** In the longer term, we recommend investigating the role of BioCommons as a "Data Broker[91]" for the international repositories to publish the collected and harmonised data. We have previously envisaged a "staging post[92]" for data and related metadata submitted from Australia to relevant international repositories, which is aligned with this idea. Our previous infrastructure roadmaps (listed in section 1) also suggested that providing in-person support from experts in formatting data and curating metadata to comply with repository format requirements similar to the previous EMBL-ABR data chaperoning service offered by QCIF. This should also be considered in the longer run if the Australian BioCommons decides to go down the path of establishing a data brokering partnership with international repositories. Consistent with our National Bioinformatics Infrastructure Roadmap plans, we recommend initially **exploring** the costs and benefits of any potential implementation of a staging post where	In addition to the training events and documentation improvements recommended in the short and medium term solutions, Australian BioCommons could also play the role of data broker to support Australian researchers curate and submit their data to international repositories. Australian BioCommons is in a unique position to actively establish collaborations and partnership with the international repositories as well as leveraging the expertise of experienced teams locally e.g. the data chaperoning team detailed in section 2.3. Acting as a data broker would enable provision of various artefacts, tools and platforms to local researchers, including a possible national staging post with incorporated guides, knowledge bases and metadata template information. This will, however, require further exploration and in depth requirement analysis particularly to understand the feasibility and costs (in both establishing and maintaining) such a staging post. Providing such service for Australian researchers could potentially resolve a number of issues indicated by the interviewees in this report spanning	Community Engagement Team (BA-led) • DevOps role to support the exploration Partners: • TBC

[91] Your tasks: Data brokering | RDMkit rdmkit.elixir-europe.org/data_brokering
[92] Nelson, TM., Griffin, P. and Christiansen, JH. Genome Annotation Infrastructure Roadmap for Australia, 2020, page 12. 10.5281/zenodo.3942716

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		researchers have access to relevant documentation, collated advice/guides and templates of repository formats required for data and metadata. If there is seen to be value, in the future, we would propose working closely with the many groups of stakeholders to produce a business case to determine the feasibility of implementing a staging post/ brokering service.	multiple themes.	
Short Medium Long	Brad The Beginner Biologist Leia The life scientist Beena The Bioinformatician Cameron The computer scientist	**R12 - Overall Awareness Raising of the activities outlined in R1-R11:** We recommend establishing an ongoing campaign, via social media (e.g. twitter, newsletters etc) or other methods to create awareness about the key topics outlined in this report eg. significance of prioritising FAIR principles during data management life cycle, the data management process/steps, international guides for data publications etc. The campaign to raise awareness should aim to demystify the topic of experimental metadata (and associated information standards) and empower researchers to publish their data in various open-access public data repositories. As a result, researcher's data will be accessible using the unique identifiers and findable with	This ongoing campaign is crucial to raise awareness about the fundamental topics related to research data management including the data publication stage. There exist relevant documentation hosted globally[93,94] that can be disseminated and promoted through suitable channels. However, this awareness is an overarching goal that will benefit from the activities recommended in short and medium term solutions. There can be webinars or workshops (as suggested in R4 and R5) dedicated to building awareness.	BioCommons team: • Communications & Training • Community Engagement Speakers/Trainers: • External - TBD

[93] Your tasks: Data publication | RDMkit rdmkit.elixir-europe.org/data_publication
[94] Data Management in Simple Steps | RDM Guide rdm.elixir-belgium.org/data_management_steps

Timeframe	Relevant User Personas catered for	Recommendation	Reasoning	Concerned Team /Individual
		well-described metadata utilising well established vocabularies and ontologies. This paradigm shift is a long term goal and can be influenced by the completion of the medium and short term goals detailed in this report. The process of data validation and submission should be made error-free, time-efficient and easy to follow in order to steer the researchers in the direction of thinking about FAIR principles.		

5. Conclusion

The Omics Data Publishing to International Repositories report provides recommendations to streamline data publication to international repositories by Australian researchers. These recommendations align with the principles of Australian BioCommons[95]. We propose to:

- provision alternative interactive submission methods for life science researchers where feasible (R1- Deploy ENA upload tool within Galaxy Australia)
- address the data submission challenges experienced by Australian researchers through collaborations with various international repository management teams to design new training programs and improved documentation to be made available via the host repositories (R2, R4, R5, R6, R7, R8, R10)
- enable researchers to more confidently prepare and submit their data to international repositories through raising awareness and the support of community building and training (R10, R12).

Our efforts will align with existing global resources wherever possible, leading to the improvement of documentation, tools, and training materials related to life science research data management and omics data publication.

[95] biocommons.org.au/about

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
V 1.0	25/11/2022	Farah Zaib Khan Jeff Christiansen (Australian BioCommons)	First draft for feedback within BioCommons Coordination Hub
V1.1	05/12/2022	Farah Zaib Khan Jeff Christiansen (Australian BioCommons)	Version for feedback from EBI repositories (ENA & Metabolights).
V1.2	20/11/2023	Farah Zaib Khan Jeff Christiansen (Australian BioCommons)	Version with incorporated feedback from EBI-ENA

Appendix 1

Table 3. User stories

Biocommons Community (relevant to this user story)	As a...	I require/need/would like...	So/because...
Genome Assembly	Researcher	to have a built-in mechanism during file generation that makes the files compliant for submissions (including correct formatting of data files and metadata) to international repositories such as NCBI/ENA.	that I can share my genome assemblies publicly via international repositories without difficulty.
Genome Assembly	Researcher	to have easier and faster ways to share my genome assembly data to international repositories such as ENA or NCBI	I find the process for uploading data to ENA or NCBI to be horrible and painful.
Comparative Genomics	Researcher	Streamlined methods/tools/processes for submission to Genbank	I find the Genbank submission process increasingly difficult. More generally, the requirements for metadata are getting out of hand, often with dozens of fields whose only realistic entries are "does not apply in my case" or "I have no idea what you mean with this field".
Microbiome Analysis	Researcher	assistance with the submission process for metagenome-assembled genomes	we always want to publish our data to international repositories but the process is convoluted
Genome Assembly	Researcher	to easily share my genome assembly data to international repositories.	uploading data to NCBI is difficult and time consuming
Genome Assembly	Researcher	tools to submit data to databases	I can publicly share my genome assemblies and the raw data
Genome Assembly	Researcher	to have a connection between NCBI/SRA and BioCommons proposed infrastructure	raw data can be sent to global repository for data sharing publicly
Genome Assembly	Researcher	To access better information/policy about diploid genomes submission to ENA/NCBI	I can share my genome assembly data to international repositories such as ENA or NCBI without the issues or errors
Genome Assembly	Researcher	to easily share my genome assembly data to international repositories.	the NCBI/SRA/GenBank submission process is daunting
Genome Assembly	Researcher	to be able to easily share my genome assembly data to international repositories such as NCBI.	NCBI takes a long time
Genome Assembly	Researcher	some assistance with the NCBI metadata requirements	that I can share my genome assemblies to international repositories without difficulty.
Genome Assembly	Researcher	to have the ability to send some raw data from the Bioplatforms Data Portal to NCBI/ENA for long term data storage	NCBI/ENA are the appropriate long term storage location for raw data
Genome Assembly	Researcher	a streamlined method for uploading raw data from Bioplatforms Data Portal to a repository	I can simplify the process
Genome Assembly	Researcher	clearer descriptions of meta-data sets for Bioplatforms and how this relates to the requirements of international repositories	the process of repository submission becomes clearer and is less onerous
Genome Annotation	Researcher	Need a way to make genome annotation files with manual annotations publicly available	NCBI do not allow the incorporation of manual curation in their genome annotation deposits

Biocommons Community (relevant to this user story)	As a...	I require/need/would like...	So/because...
Genome Assembly	Researcher	to have a built-in mechanism during file generation that makes the files compliant for submissions (including correct formatting of data files and metadata) to international repositories such as NCBI/ENA.	that I can share my genome assemblies publicly via international repositories without difficulty.
Genome Assembly	Researcher	to have easier and faster ways to share my genome assembly data to international repositories such as ENA or NCBI	I find the process for uploading data to ENA or NCBI to be horrible and painful.
Comparative Genomics	Researcher	Streamlined methods/tools/processes for submission to Genbank	I find the Genbank submission process increasingly difficult. More generally, the requirements for metadata are getting out of hand, often with dozens of fields whose only realistic entries are "does not apply in my case" or "I have no idea what you mean with this field".
Genome Assembly	Researcher	to have access to appropriate training resources on how to link genomic data into other species occurrence registries/data resources	that I can share reference genomes
Genome Assembly	Researcher	to have access to appropriate training resources on how to publish phylogenomic data to national and international species taxonomies/directories.	to perform, store and share phylogenomics data
Genome Assembly	Researcher	to have access to appropriate training resources on how to share and link conservation data with conservation organisations and other species registries/data resources.	that I can analyse, store and share conservation genomics and phylogenomics data

Appendix 2

Table 4. Questionnaire

Themes	Follow-up Qs	Additional notes
Data type	What type(s) of data do you analyse or is generated as a result of your analysis/experiment?	
Scale of data to be submitted	Have you encountered issues when you tried scaling up the data to be submitted?	
When is the data submitted	When in the research life cycle? Why at that particular stage?	How will things like metadata capture or data preparation change if you submitted data early and often?
Pre-submission storage	What is your mode of storage for data prior to submission? Why did you choose this mode of storage?	How easy was it to move data around from the pre-submission storage to the repository of your choice? What methods did you use to transfer data (ftp?)
Level of expertise	How would you describe your level of experience with CommandLine Interface (CLI) and Application Programming Interface (API)?	
Contextual info - about samples, experiment, people involved -- biocuration		
	At what stage of the research lifecycle do you begin recording contextual information/metadata? Why at this stage?	At what stage of the research lifecycle would you like to get an identifier for studies and samples?
mode of documentation	How do you keep record of the contextual information/metadata for a given project?	If the metadata is not stored using template spreadsheets, what is your workflow to transform the metadata from your format of recording to repo-acceptable format?
metadata storage	Where do you store the contextual information/metadata?	
controlled vocabularies	Do you use controlled vocabularies or ontologies to represent the contextual information?	(e.g. Darwin Core (DC), GO, NCBI Taxonomy, MeSH, EFO)
Vocabulary lookup services	Do you use any vocabulary lookup services to help you complete the contextual information/metadata	(e.g. NCBI taxonomy browser, OMIM, OBO Foundry, ZOOMA, BioPortal etc.)
checklists	A checklist defines a set of mandatory and recommended contextual information values for a given type of data (e.g. ERC000012). Do you use/conform to any standardised sample checklist provided by the repositories you plan to submit your data to? How was your experience with these?	
guidelines	There exist well-established guidelines that outline the minimum contextual information/metadata that should be included when describing a study (e.g. MIAME (Minimum Information About a Microarray Experiment), MINSEQE (Minimum Information About a Next-generation Sequencing Experiment). Do you refer to such guidelines when keeping track of the contextual information/metadata related to your project?	Are there templates (spreadsheets/JSON files) available to be downloaded and filled in?

Themes	Follow-up Qs	Additional notes
data model	A data model provided by the host repository guides you about what parts of your research project can be represented by which metadata objects and this determines what you need to submit. The data model helps to understand how different elements e.g. samples, experiments and studies are related to each other in a repository. Do you organise your data and its contextual information/metadata using a data model provided by the repository where you intend to submit your data?	(for e.g. ISA data model, ENA data model)
Submission to international repositories		
preference	Generally, do you have a preference for the international public repository to submit your data?	Which one? Why this one?
Alternative	Would you consider choosing an alternative repo?	If not, why not? What issues need to be addressed to make the alternative work?
mode of submission	What mode of submission do you prefer/comfortable with?	What is the reason behind it?
time commitment	On average, how long does it normally take to submit your data to the repository of your choice?	
submission workflow	In a few sentences, describe your current workflow/process in place for submitting your data to the repository of your choice.	
External Support	Have you ever used any additional software (e.g. Galaxy or COPO) as a top layer that facilitates and brokers the submission of the data to the repository of your choice?	How did it improve your experience of submission?
Challenges		
Difficulty level	Overall, How hard was the process of data submission and publication? if you have to rate from 1 being extremely easy and 5 being extremely difficult	
Clear information in literature/repositories about where to submit	Do you have clear instructions/information about where and how should you submit your data?	
Appropriate training provided by either the hosting repositories or independent organisations/3rd party initiatives	Do you have access to in person/online training material from any source detailing the data submission process and to inform your choice of the repository?	If not, can you please elaborate which phase of data submission do you require training for and what mode of training would you prefer (e.g. webinar, workshop, one-to-one training)?
Ease of use	How easy/difficult did you find the interface/method of uploading/submission in terms of user experience?	
Clear information about the contextual information/metadata requirements	Does such information exist? If yes, how easy/difficult was it to find it?	if no, how do you think we should better present/share this information with researchers
Technical difficulties	While submitting your data, have you encountered technical difficulties and given up in the middle of the process of submitting?	Can you please describe the nature of the issues and the measures you took to resolve these issues?

Themes	Follow-up Qs	Additional notes
Incomplete information record	Have you encountered a situation where you had not captured/recorded the required metadata during the course of your research project and it hindered you to submit your data to the repository of your choice? If yes, please share your experiences	
Assistance from repository support team	Have you received assistance/support from the helpdesk/support team of a particular repository when required?	If not, did you get the extremely delayed response that it was not relevant anymore?
Other challenges	if not covered above	
Interviewee's suggestions/insights	Overall, how do you suggest improving the data submission/publication process for the community? What can biocommons do to help the researchers in the process?	

Appendix 3

The participants in our community engagement activities can be categorised into the following four user personas (See Figure 3).

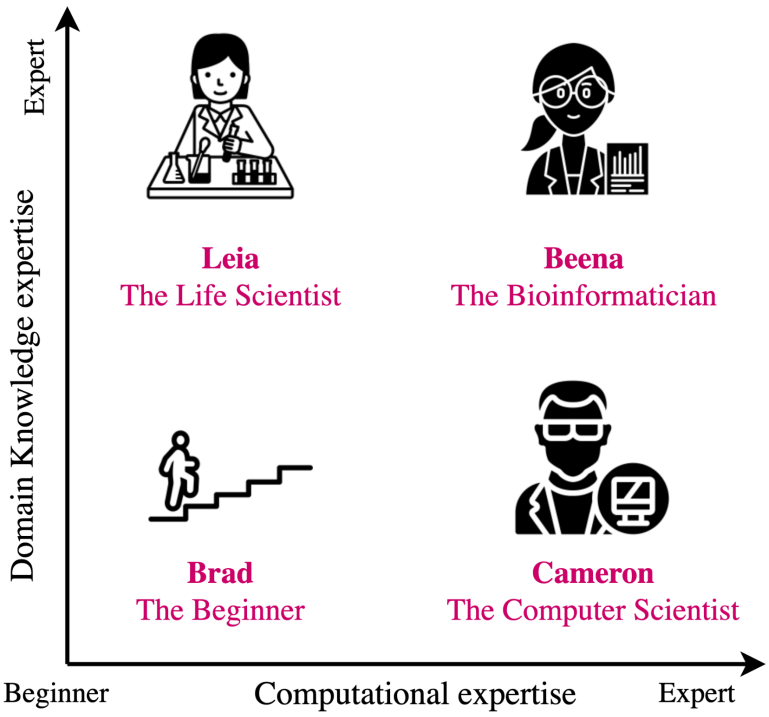

Figure 3. Sample user personas to categorise the target audience of the solution space proposed in this report.

– Leia, The Life Scientist:

> Leia is a life scientist who is currently employed as a post-doctoral fellow in a group that works on microbial ecology. With a background in biochemistry and ecology and several years of experience in different labs, Leia is an expert in designing and performing wet lab experiments. Leia analyses the data produced in the lab she works in and also interacts with the bioinformaticians in her lab to design analysis methods for her data.
>
> In the analysis process, Leia occasionally utilises bioinformatics tools and services. In order to support her research. Leia also interacts with several online data repositories (including those hosted by EMBL-EBI and NCBI) to download supporting datasets or to make her own data available at the time of research publication. She is comfortable with user friendly GUIs to carry out any analysis or data handling. She needs additional support if advanced computing skills are required to complete a task.

– Beena, The Bioinformatician:

> Beena is a bioinformatician with a background in mathematics and statistics. She has extensive experience in building tools and services for several projects.

Currently, Beena is working on several projects where she is developing in-house scripts and pipelines to analyse the functional genomics data produced in the lab she is associated with. Beena is not involved directly in wet lab experiments but frequently extracts relevant test datasets from public resources for the pipeline evaluation. Beena is very comfortable dealing with data and methods via the command line interface. At the time of publishing her research and assisting her lab fellows in publishing their data, she finds it time consuming and difficult to glean relevant information about submission guidelines and requirements of the relevant international repositories.

– *Cameron, The Computer Scientist:*

Cameron has a background in computer science with a Bachelor's degree in computer science and Master's degree in data science and statistics. Cameron has extensive experience and training in applied mathematics, statistics and computing. Recently, Cameron has started working on life science projects to apply his expertise in statistics and computing. Cameron is a beginner, only now learning about the life science domain as part of his job. Hence, Cameron frequently consults the life scientists he works with for domain specific discussions.

Currently Cameron has joined an organisation carrying out research in plant pathology where he is interested in developing diagnostic tools to carry out genome-based surveillance of fungal crop pathogens. With his background, Cameron is an expert dealing with the data and methods via command line interfaces, as well as via the GUI. However, Cameron needs additional support from his colleagues and through self-learning to understand the in-depth concepts associated with transcriptomics, phylogenetic analysis and comparative genomics. In addition, Cameron is the focal person in making the submissions of the project-related data to the relevant international repositories on behalf of his colleagues and collaborators.

– *Brad, The Beginner Biologist:*

Brad is a first year PhD student with an undergraduate training in molecular biology. He has recently completed his Master's degree in molecular biology and is now enrolled in a PhD program with a group focused on conducting research exploring the factors associated with the lack of susceptibility of poultry to SARS-CoV-2. Brad will be learning theoretical concepts related to the topic of research and the latest wet lab techniques for carrying out his experiments (which will include high-throughput nucleic acid sequencing). In addition, Brad is also a novice in computational skills e.g. he has a limited experience interacting with data or computational methods, or tools via either a GUI or the command line interface.