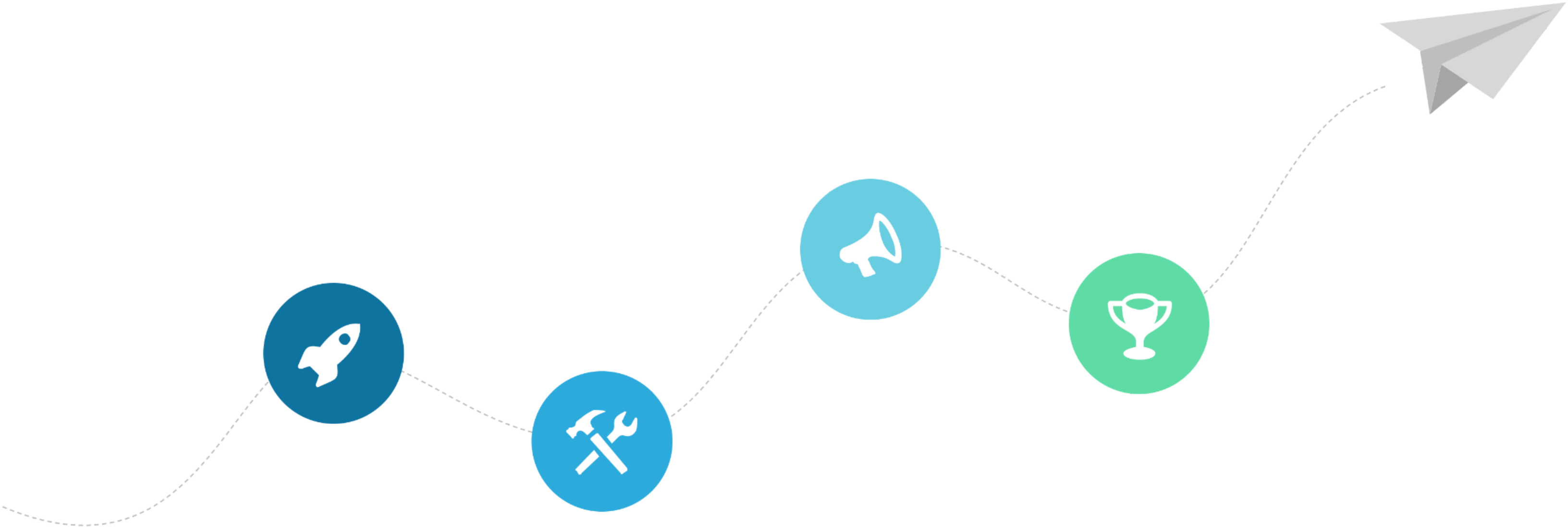
HUMAN GENOMES PLATFORM PROJECT

Data and Metadata Archiving

FEASIBILITY REPORT

Dec 2023



Authors

in alphabetical order by surname

Copty, Joseph - Garvan

Cowley, Mark - ZERO

Evans, Ben - NCI

Hoffman, Oliver - UMCCR

Holliday, Jessica - BioCommons

Kaplan, Warren - NCI

Koufariotis, Ross - QIMRB

Kummerfeld, Sarah - Garvan

Pope, Bernard - BioCommons

Reisinger, Florian - UMCCR

Robinson, Andrew - NCI

Shadbolt, Marion - BioCommons

Syed, Mustafa - ZERO

Wong-Erasmus, Marie - ZERO

Acknowledgements

The HGPP formed part of Australian BioCommons' Human Genome Informatics initiative and was funded by NCRIS via the Australian Research Data Commons (https://doi.org/10.47486/PL032) and Bioplatforms Australia. Contributions were also made by partner organisations: Australian Access Federation, Garvan Institute for Medical Research, National Computational Infrastructure, QIMR Berghofer Medical Research Institute, The University of Melbourne Centre for Cancer Research, the ZERO Childhood Cancer Program and Children's Cancer Institute.

Table of Contents

Glossary

CEGA	Central EGA. A a controlled-access repository for the long-term storage of human data. The central and original node of the EGA co-managed by CRG in Barcelona and the EBI in the UK.
Controlled access	Where a data applicant must go through an application process with a data access committee to access data in accordance with a data access policy.
CRG	Centre for Genomic Regulation, based in Barcelona, Spain.
CSV	Comma separated value files. A format to specify tabular data in a plain text file.
DAC	Data Access Committee. A group that approves or rejects applications to access controlled data, may be the data custodians or institutional representatives working on their behalf.
dbGaP	Database of Genotypes and Phenotypes. A US-based managed access repository for sensitive human data and metadata related to studies primarily funded by the NIH.
DPO	Data Protection Officer.
DUOS	Data Use Oversight System. A software package to manage data access requests.
EBI	European Bioinformatics Institute. Based at the Wellcome Genome Campus, Hinxton, UK.
ELSI / ELSA	Ethical, Legal and Social Implications/Aspects of research.
FEGA	Federated EGA. A distributed node of the EGA located within a country that connects to the Central EGA.
FHD	ELIXIR Federated Human Data community.
FHIR	Fast Healthcare Interoperability Resources. A standard for exchanging digital healthcare information between systems.
FTE	Full-time equivalent.
GSA-Human	Genome Sequence Archive for Humans. A controlled access repository run by the China National Centre for Bioinformatics and National Genomics Data Center (CNCB-NGDC).
JGA	Japanese Genome Phenome Archive. A controlled access repository run by the National Bioscience Database Centre (NBDC) and Japan Science and Technology Agency (JST)
LocalEGA (LEGA)	Software deployed by federated nodes of the EGA to enable non-sensitive information exchange between the node and the central EGA.
[meta]data	Data and its associate metadata.

NCBI National Centre for Biotechnology Information (USA). Offers a number of genomic data sharing resources, including SRA and dbGaP, and is based out of the NIH National Library of Medicine.

NeIC Nordic e-Infrastructure Collaboration.

NIH National Institutes of Health. Based in the USA.

OMOP CDM Observational Medical Outcomes Partnership Common Data Model. A standard for the structure and content of observational data, mainly aimed at Electronic Medical Record use cases.

REMS Resource Entitlement Management System. A software platform for managing and streamlining the Data Access Request and approvals process.

SDA Sensitive Data Archive.

1. Background

The generation of human genomics data is growing at an unprecedented rate [1]. The retention and sharing of this data is important for a number of reasons, including:

- Ensuring transparency and reproducibility of results.
- Validation of new discoveries in comparable datasets.
- Ability to reanalyse data with new methods and gain new insights.
- Re-use of data for purposes not imagined by original data collectors.
- Integration of datasets to gain more statistical power.

The availability of human genomics data for integration and reanalysis has the potential to contribute to improved health outcomes through biomarker discovery, clinical trials, and precision medicine. In addition, generating sequencing data for research is expensive and largely funded through government grants. Therefore, deriving the maximum value from these data and ensuring it benefits the wider community is paramount.

As a result, within the human genomics community there is momentum towards making genomics data Findable, Accessible, Interoperable and Reusable (FAIR) [2–7]. Increasingly, many leading scientific journals require human genomics data to be submitted to a recognised archive before publication [8–10]. In addition, the National Institutes of Health (NIH), which funds the vast majority of all health-related research in the US, mandates that all human genomic data they fund is shared [11]. This has led to large amounts of human genomic data being submitted to global archives in order to meet these requirements. The value and utility of these data depend not only on the data being findable but also on the quality of the contextual metadata describing it. Typically, archives place minimal requirements on contextual metadata that may describe the clinical, demographic or phenotypic aspects of the samples, as well as information about the experiment/data acquisition methods used. Limiting these requirements can make the submission process easier but restricts how the data can be found and/or reused by others [12].

In Australia, there are no national mandates or standards for how human genomics research data should be managed, shared, and preserved long-term. Peak funding bodies, the National Health and Medical Research Council (NHMRC) and the Australian Research Council (ARC), give general guidance and recommendations, but these are not tailored to the human genomics domain [13,14]. Guidance is aimed mainly at institutes to implement clear policies around ownership, management, preservation and accessibility of data [13,14].

The Australian government has recognised the need for a national approach to managing human genomics data (NAGIM) [15,16]. A blueprint for how this infrastructure could be established as well as a technology piloting phase have been undertaken [16,17]. Australian human genomics research data is often siloed within institutes, or housed in difficult to access overseas archives, limiting the benefits that could be derived from these valuable data to Australians.

The Human Genomes Platform Project (hereafter referred to as ‘the HGPP’ or ‘the project’) is a collaborative research project aiming to overcome some of the challenges and limitations outlined above. Its goal is to enhance secure and responsible human genomic data sharing for research purposes. The project partners represent many of the largest human genome sequencing and analysis organisations in Australia.

The goals of the data and metadata archiving sub-project within the HGPP are to:

- Understand the needs of stakeholders when submitting, downloading, and managing datasets in international repositories and the challenges they have faced.
- Investigate the options and requirements for establishing national human genomics repositories in Australia.
- Provide technical insight into implementation options.
- Link to international communities, platforms, standards, and solutions.

Human Genomic [meta]data archiving

Due to consent, ethics and Australian privacy laws, raw human genomics data must be submitted to a controlled access repository [18,19]. Several controlled-access repositories exist offshore for the secure archiving of human genomic data and metadata. These may house data derived from Australian-based individuals or studies, where consent and ethics procedures allow (see Table 1 below).

The controlled access repositories considered for this report were Federated EGA (FEGA), database of Genotypes and Phenotypes (dbGaP) and Terra Data Repository (TDR). The FEGA model was discussed in detail in the Background section, below we compare and contrast dbGaP and TDR.

Appendix D outlines the different features or requirements for a platform to handle genomic and data and metadata archiving.

Table 1. ***Summary of controlled access repositories for human genome data available to Australian researchers.***

Repository	Publication	Location	Data Access model	Restrictions	Count of Aust. datasets
European Genome-Phenome Archive (Central EGA)	Freeberg et al. [20]	UK & Spain	Data access Managed by DAC of each dataset	Open for anyone to submit	139[1]
Database of Genotypes and Phenotypes (dbGaP)	Tryka et al. [21]	USA	Managed centrally by dbGaP committee	NIH-funded research automatically accepted, others on case-by-case basis	48[2]
Japanese Genome-Phenome Archive (JGA)	Okido et al. [22]	Japan	Data access managed centrally by the JST-NBDC	Submission needs to be approved for submission by the JST-NBDC	0[2]
Genome Sequence Archive for Human (GSA for Human)	CNCB-NGDC Members and Partners [23]	China	Data Access managed by the custodian/DAC of each dataset	No apparent restrictions but mainly holds data from China	0[2]
Terra Data Repository (TDR)	Data Science Platform at the Broad Institute	Dependent on cloud region (default region is USA “us-central”)	Data Access managed by the custodian/DAC of each dataset	Open for anyone to submit	0[2]

dbGaP

The dbGaP is managed by the National Center for Biotechnology Information (NCBI, Bethesda, USA) and primarily holds data generated by projects funded through US NIH grants. Data from other sources may

[1] Queried from publicly available EGA metadata as at 15/03/2023

[2] Estimated from ’Australia’ keyword search

be submitted, but must go through an approval process [24]. At the time of writing, 48 NIH-funded datasets containing Australian data have been submitted to dbGaP from a range of NIH institutes (see fig. 1). The US National Cancer Institute (NCI) provided funding for 18 of the datasets while the National Institute of Neurological Disorders and Stroke (NINDS) funded a further eight datasets, with a further 12 institutes contributing to the remaining 22 datasets (Figure 1).

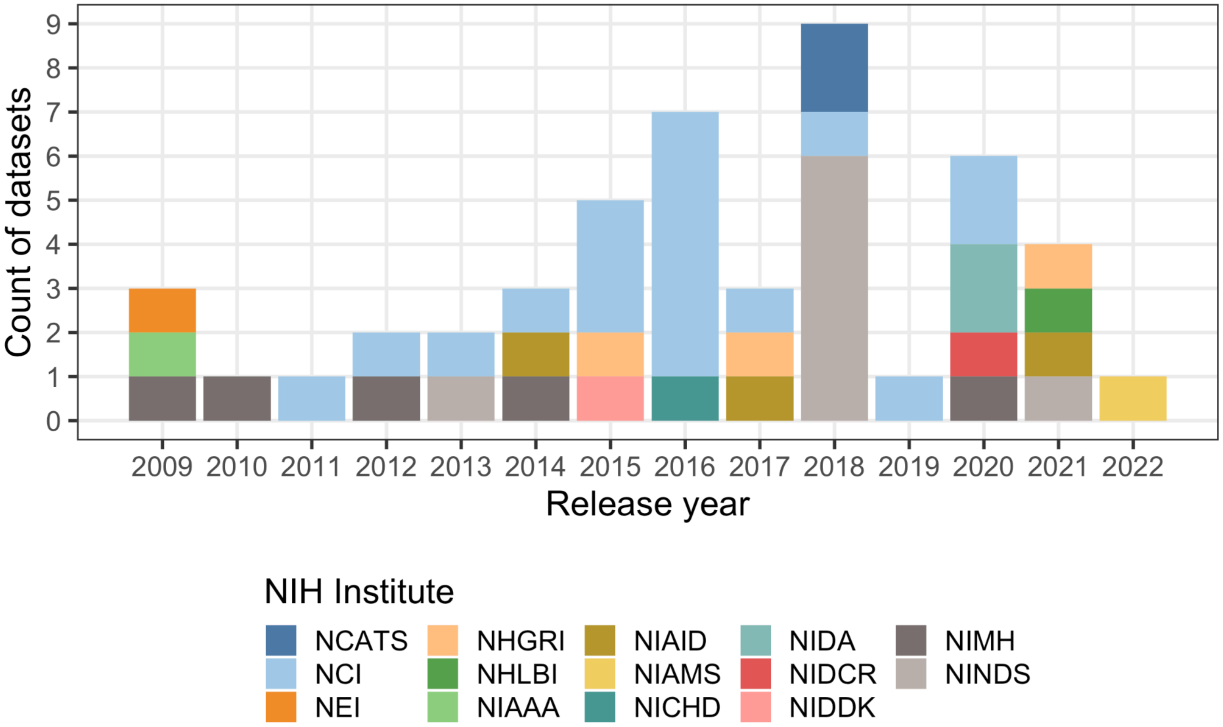


Figure 1. ***Count of datasets with "Australia" mentioned in study description submitted to dbGaP over time. See NIH List of institutes for initials used above.***

The dbGap platform offers free storage and a stable persistent identifier for the submitted dataset which may be used for publications relevant to the dataset. The platform is popular in the human 'omics domain and provides researchers controlled access to data sets. Access to submitted data is restricted to NIH investigators. That said, there is also a large international user base recorded on their website with 192 Australian PIs registered and 1388 requests to access data made (dbGaP summary statistics).

In addition, dbGap offers the ability to:

- Upload de-identified data using secure data transfer protocols.
- Provide a digital object identifier (DOI) for publications.
- Search and browse the data collections on the platforms via a catalogue.
- Apply for datasets with restricted access control.
- Share the datasets using controlled access.
- Use as an archive of datasets with no compute connected.

The dbGaP platform hosts the data on their storage servers. As well as storing data itself, dbGaP also allows data to be uploaded to cloud data storage and managed by an external data source from a 'trusted partner', such as through NIH-hosted repositories. 'Trusted Partners' are established through a contract between an NIH funding agency and an organisation that is able to meet core NIH standards for data quality and management [25]. The Kids First Data Resource Portal is an example of this, where subject and sample IDs are maintained in the dbGaP but data access and download is managed directly in the Kids First Data Portal [26–28].

Terra Data Repository (TDR)

The Terra Data Repository is a feature of the Terra cloud platform, built and operated by the Broad Institute. Terra is a scalable platform aimed at solving some of the pain points for genomic data research, including data sharing, storage, and analysis.

It is important to point out one key difference when using TDR compared to other candidate repository solutions: TDR allows one to create datasets for archiving and sharing data at either the data custodian or the data consumer's expense. The cost for storing data and sharing it will incur cloud storage and egress fees. Terra data can be hosted on Google cloud platform (GCP), as well as the newly-introduced compatibility with Microsoft Azure cloud (at the time of writing this report, a "public preview release" was available to allow early access to customers who apply [29]).

Some features to highlight include:

- Upload any type of data to TDR via GCP or Microsoft Azure data bucket. You can also upload data via the Terra platform interface. You may choose to store the data in a GCP bucket hosted on Australian nodes. (TBC for Azure)
- Share the data collections using controlled access.
- Browse existing datasets and apply for access after logging into Terra.
- Secure storage. Security is built into the Terra platform, however, it may cause some vulnerability as the security is enforced via user-defined permissions and roles.
- Terra uses a cloud centric approach. This provides the ability to perform computations on the data using various tools, such as WDL pipelines, Jupyter notebooks, RStudio or Galaxy, which means there is no need to download the dataset.
- Allow the sharing of a subset of data using concepts called "Assets" and "Data Snapshots". This allows highly customised and versioned sharing of data, giving flexibility to both the consumer to only share what they want to and for the consumer to only access what they need.
- No limit to data sizes of a cohort or submission (cost will increase with more data).
- Define your own schema and add custom metadata. There are some recommendations available in the user documentation.

However, TDR lacks some features that are common for the other platforms considered, including:

- Does not have fully-subsidised storage. Data owners are expected to pay for cloud storage. Additional to the storage costs, egress costs will also be charged to the data owner or data consumer based on the bucket setup.
- It is unclear at the time of writing if Terra offers a service for minting a DOI for each dataset that can be used for publications. However, they do offer Unique IDs (UUIDs) for datasets, which may be used as an alternative. It is not clear whether these UUIDs would be accepted by publications as proof of data deposition. Further investigation is needed to understand whether UUIDs meet publication and funding requirements for proof of data deposition.

EGA

The European Genome-Phenome Archive (EGA) is a controlled access repository for data that has specific data use and permissions [20]. It is managed by Elixir and is one of the only repositories of its type available for submissions from Australian researchers. It offers many favourable features, including the ability to:

- Upload encrypted deidentified data.
- Provide a digital object identifier (DOI) for research publications.
- Search and browse the data collections on the platform via a catalogue.

However, researchers can face significant challenges with the current processes for submission [30,31]. The partners involved in this project have experience with submitting to the EGA, their workflows are summarised in appendix A. Some of the major challenges faced by submitters when trying to submit include:

- Limited upload/download speeds from Australia to Europe (the AARNet link up through Singapore and into Europe is 10Gbit/s).
- Limited capacity of staging areas for user submission data files (<10TB).
- Helpdesk accessibility due to time zone (~9-11 hours difference to Australia).
- Difficulty understanding metadata and submission requirements.
- Requirements for data and metadata to remain within Australia under certain circumstances (e.g., where the consent, ethics, or Australian law requires it).

Specific needs and challenges are discussed in further detail in the Community Needs Analysis.

Despite these challenges, as at 15 March 2023, Australian researchers have collectively submitted over 325 TB of data to the EGA spread across 139 datasets. Data custodians represent 24 medical research institutes, universities and government agencies (Figure 2). While over 50% of datasets currently submitted are less than one terabyte in size, larger datasets, such as the Medical Genome Reference Bank [32], range up to ~80 TB.

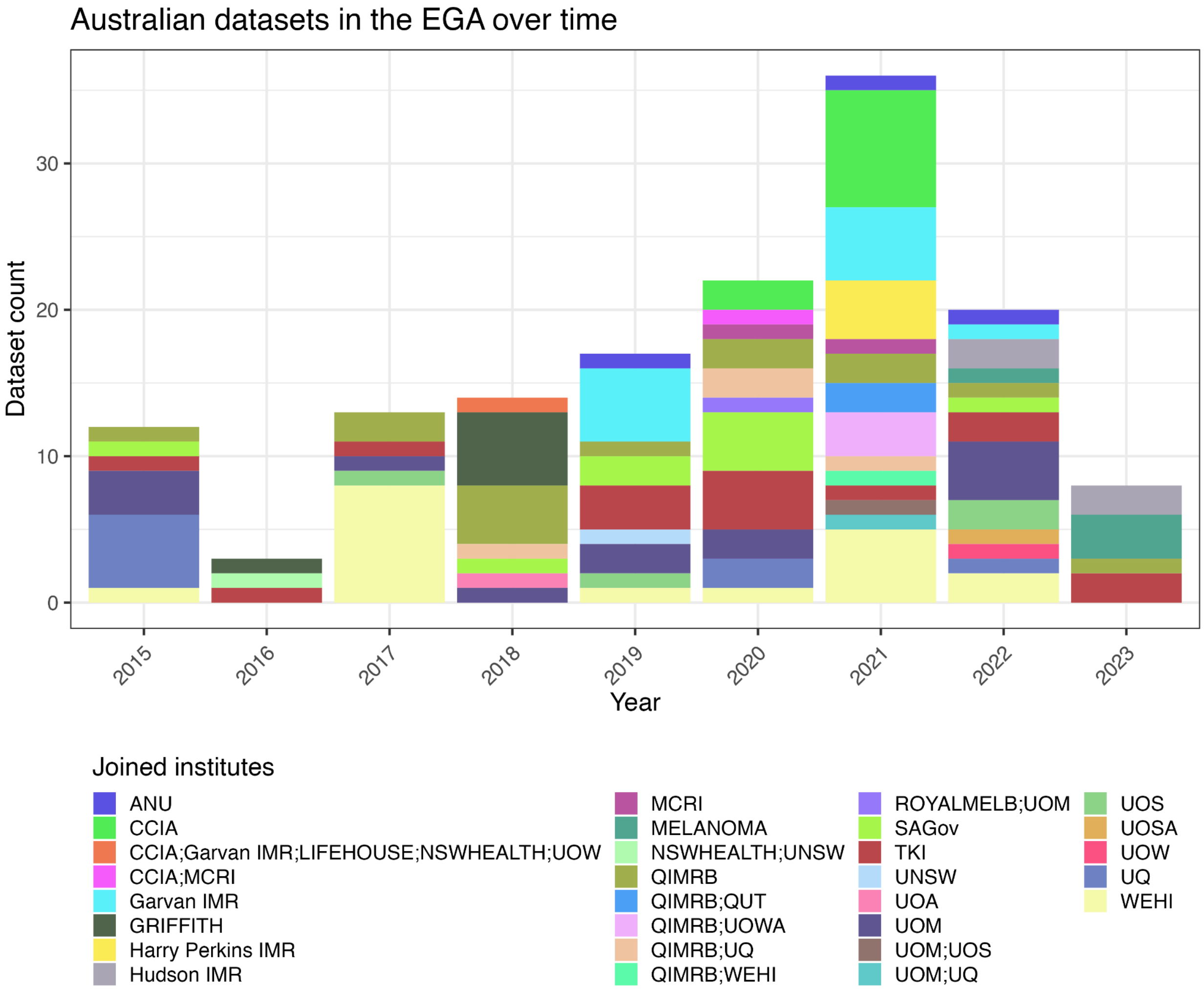


Figure 2. ***Count of datasets submitted to the EGA over time by Australian Institutes over time (data accessed May, 2023).*** ANU: Australian National University, CCIA: Children's Cancer Institute Australia, Garvan IMR: Garvan Medical Research Institute, GRIFFITH: Griffith University, Hudson IMR: Hudson Institute of Medical Research, LIFEHOUSE: Chris O'Brien Lifehouse, MCRI: Murdoch Children's Research Institute, MELANOMA: Melanoma Institute Australia, NSWHEALTH: New South Wales Department of Health, Harry Perkins IMR: Harry Perkins Institute of Medical Research, QIMRB: Queensland Institute of Medical Research Berghofer, QUT: Queensland University of Technology, ROYALMELB: Royal Melbourne Hospital, TKI: Telethon Kids Institute, UOA: University of Adelaide, UOM: University of Melbourne, UNSW: University of New South Wales, UQ: University of Queensland, UOS: University of Sydney, UOSA: University of South Australia, UOWA: University of Western Australia, UOW: University of Wollongong.

To date, the majority of controlled access Australian human genomics data is submitted to the Central EGA. However, it is likely that a large amount of data is never submitted to international archives due to the lack of a national mandate, the challenges that Australian researchers face with submission, and the legal and ethical issues associated with storing Australian human genomics data off-shore. Additionally, if the data is required for reanalysis, it is challenging to download the data form international archives. These limitations are an impediment to maximising value through discovery, integration, and reuse of data being generated by Australian researchers that is largely publicly funded. The establishment of a controlled access repository within Australia that is part of a large global network could alleviate many of these issues, with the Federated EGA presenting a potential solution.

Overview of the Federated EGA

Federated EGA Vision:

> “*Federated EGA strives to support the discovery of and secure access to human data globally, while respecting national data protection regulations, with the goal of accelerating disease research and understanding and improving human health*.”[33]

The Federated EGA (FEGA) model enables countries to establish their own national human genomics data repository, where sensitive data files remain within that country, and public metadata is stored and indexed in the Central EGA. The main driver of this shift is a change in country-specific legal regulations that mandate that sensitive human data must remain within the jurisdiction of the country in which it was generated.

Datasets have the benefits of a recognised EGA accession, discoverability within the central data portal, and compatibility with the EGA metadata standard. FEGA nodes leverage knowledge, experience, guidance, and support from the Central EGA who have been running their global repository for 15 years. Each FEGA node needs to communicate with the Central EGA in order to exchange information, such as globally unique accessions and non-sensitive metadata. This is done through the international standard Advanced Message Queueing Protocol (AMQP) [34]. Both the reference implementation, LocalEGA, and the NeIC SDA implementation use RabbitMQ to broker messages using this protocol.

The Central EGA is not prescriptive about the various other services that are built around and on top of it. Each node largely takes its own approach on the type of technology stack, data access model, and various other services that are offered alongside submission of data and metadata, and accession assignment. Country-specific approaches are discussed in more detail below. Technical aspects of the LocalEGA software options are discussed in the Candidate Solutions section.

More than 20 countries are currently engaged with the FEGA process at various levels of maturity, both within and outside Europe (Figure 3). Four nodes (Norway, Finland, Sweden and Spain) have performed end-to-end testing.

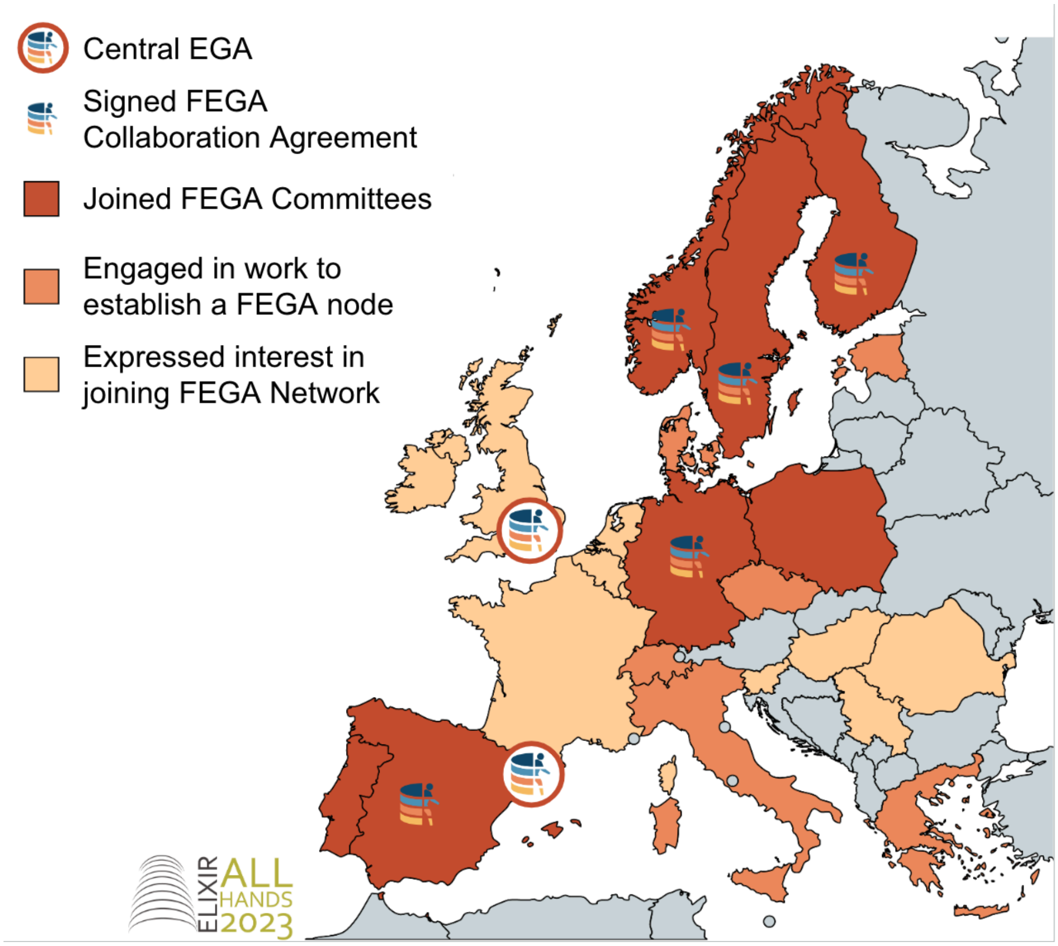
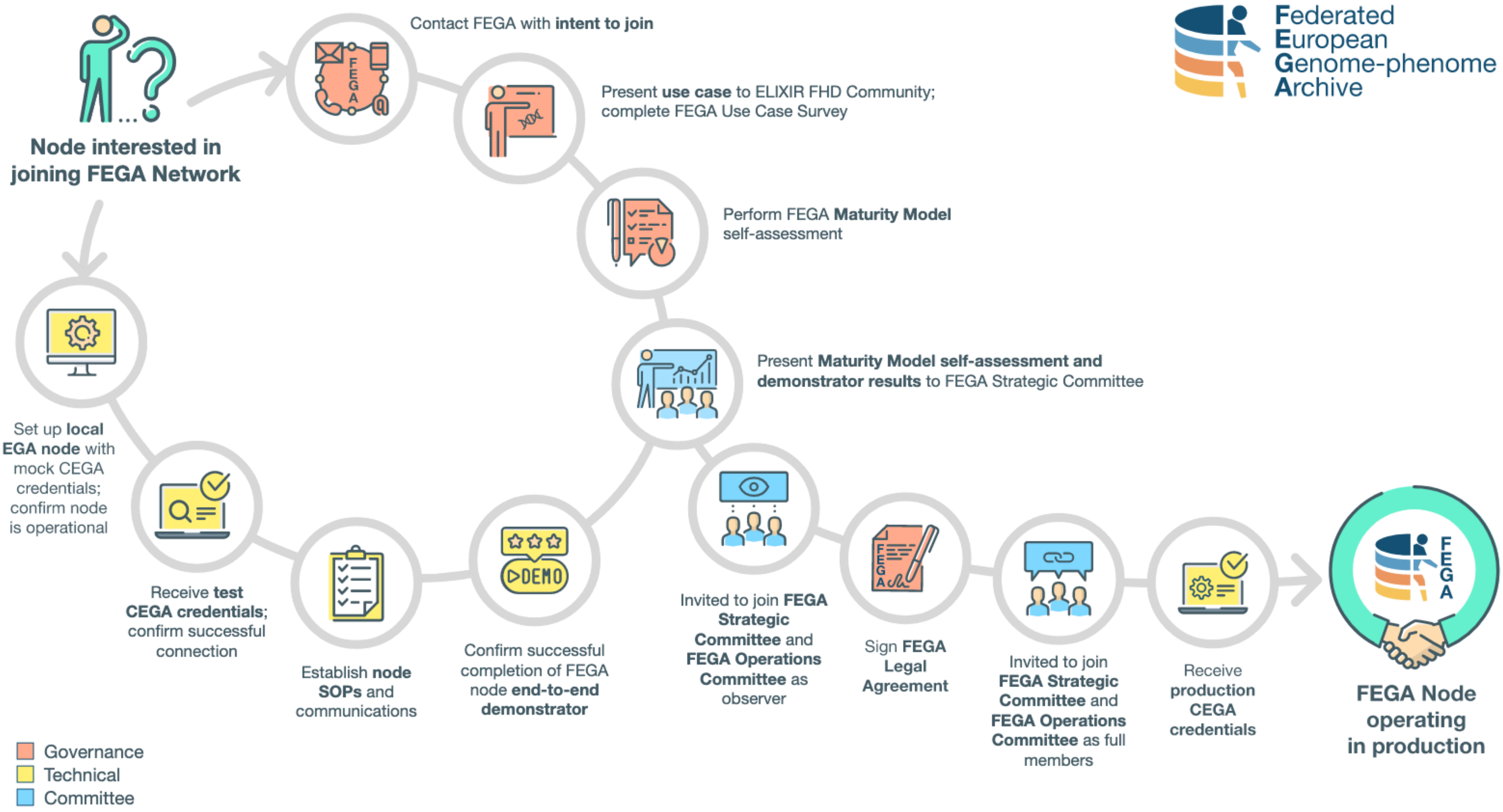

Central EGA
Signed FEGA Collaboration Agreement
Joined FEGA Committees
Engaged in work to establish a FEGA node
Expressed interest in joining FEGA Network
ELIXIR ALL HANDS 2023

Figure 3. ***Map of currently engaged FEGA nodes in Europe, also engaged with countries outside Europe such as Argentina and Canada (Source:*** [35] ***reproduced with permission from the authors).***

As the FEGA network grows, the Central EGA is working to ensure there is adequate support for prospective members to understand what is involved and navigate the path towards becoming a full member (see Figure 4). The resourcing required to establish a node spans six key areas: Governance, Legal, Data & Metadata, Infrastructure, Operations and Community & Outreach [36]. Documentation and guidelines about the resourcing required is available on the FEGA website and are under regular revision and development. In addition, the ELIXIR Federated Human Data (FHD) Community acts as a meeting hub for prospective and established nodes to exchange information and discuss progress and challenges in FEGA node establishment.



Federated European Genome-phenome Archive
Node interested in joining FEGA Network
Contact FEGA with **intent to join**
Present **use case** to ELIXIR FHD Community; complete FEGA Use Case Survey
Perform FEGA **Maturity Model** self-assessment
Present **Maturity Model self-assessment and demonstrator results** to FEGA Strategic Committee
Set up **local EGA node** with mock CEGA credentials; confirm node is operational
Receive **test CEGA credentials**; confirm successful connection
Establish **node SOPs** and communications
Confirm successful completion of FEGA node **end-to-end demonstrator**
Invited to join **FEGA Strategic Committee** and **FEGA Operations Committee** as observer
Sign **FEGA Legal Agreement**
Invited to join **FEGA Strategic Committee** and **FEGA Operations Committee** as full members
Receive **production CEGA credentials**
FEGA Node operating in production
FEGA
DEMO
🟥 Governance
🟨 Technical
🟦 Committee

Figure 4. *Path to joining the FEGA network (Source: [35], reproduced with permission from the authors).*

The draft FEGA maturity model provides a detailed overview of all the aspects involved in establishing a FEGA node and developing it towards a production instance [37]. The model is divided into domains, subdomains, and indicators, which are measured in maturity levels from 1 to 5 in order of increasing maturity. The 36 indicators are in turn classed as Essential (7), Important (14) and Useful (15) [38]. As nodes develop towards maturity, they are able to conduct self-assessment against the maturity model to determine their readiness to become a production-level instance. It is expected that production-level instances will achieve a minimum score of 4 out of 5 for *essential* indicators, 3 for *important* indicators, and at least 2 for all *useful* indicators [38]. A recent first-round self-assessment by the six most mature nodes and the Central EGA was a valuable process that enabled nodes to understand their progress and identify where improvements were required [38].

Three nodes have officially signed the FEGA Collaboration Agreement (Norway, Sweden, and Germany). This enables them to join the FEGA Strategic and Operations Committees as full members and gain access to credentials that provide production-level access to the Central EGA. Poland, Spain, and Finland are expected to be the next nodes to reach this level. The collaboration agreement was developed in consultation with the inaugural nodes and represents an agreement between the two institutes that run Central EGA (EBI and CRG) and the node-hosting institute [39].

Sweden

The Swedish node, based out of National Bioinformatics Infrastructure Sweden (NBIS)[40], currently has a sensitive data archive (SDA). However, they are working to integrate this with the FEGA to improve data discoverability through having studies searchable via the Central EGA portal. They are using their own Swedish Research Cloud and a basic command line submission interface [41] to ingest data into the archive. They also help to form tighter links between the SDA and their secure human data compute resource, Bianca. They have conducted a full end-to-end demonstrator and recently signed the FEGA collaboration agreement[42].

Finland

The Finnish FEGA node is hosted by the CSC - IT centre for science. The LocalEGA software is housed within existing Finnish infrastructure that is already set up for secure human data processing and will form the means of data archiving as part of their Secure Data Services. The Finnish node is not yet in production but they have signed the FEGA collaboration agreement [43]. They have completed a successful end-to-end test and have comprehensive user documentation and informative videos available [44].

Germany (GHGA)

Germany is establishing the German Human Genome-Phenome Archive (GHGA), which will incorporate a Federated EGA node alongside standardised data analysis, data visualisation, integration, and cloud-based analytics (see Figure 6). The program of work is funded through National Research Data Infrastructure (NFDI e.V.) via the German Research Foundation (DFG).

Human genomics data management in Germany is highly distributed within the country leading to a largely scattered approach [45]. This presents challenges for the establishment of FEGA infrastructure. They are aiming to overcome this by establishing six data hubs in key locations around the country. The overall project has been divided into eight distinct workstreams: ELSI, Architecture, Workflows, Project Management, Metadata, Outreach, Training, and Data Hub Operations. These are explained in more detail on their website.

GHGA have set up their own metadata catalogue to improve discoverability of datasets that are managed by their partner institutes. The catalogue is based on their own metadata model, which provides more detailed information while remaining compatible with the EGA metadata model[46]. Access to data is currently a manual process through a template email to the data owner.

Spain

The Spanish node of the Federated EGA is set to become one of the six inaugural nodes. At the time of writing this report, they are in a testing phase and have established a workflow for data and metadata submission, as well as data delivery [47].

Norway

The Norway FEGA node (NFEGA) is hosted by the University of Oslo [48]. The NFEGA node forms their secure data archive deployed inside their secure infrastructure for sensitive data called TSD (Tjenester for Sensitive Data) [49]. They use the Life Science AAI for authentication for services to submit and retrieve data files as well as Central EGA credentials for metadata submission [50]. NFEGA have developed a tool to assist with data import and export called lega-commander [51], which securely submits encrypted data files into their archive.

2. Australian Community Needs Analysis

To determine Australian community needs for a national controlled access data repository, we conducted a workshop with representatives from our project partners: The University of Melbourne Centre for Cancer Research (UMCCR), Queensland Institute of Medical Research Berghofer (QIMRB), the ZERO Childhood Cancer Initiative based at the Children's Cancer Institute (CCI), the Garvan Institute for Medical Research, and the National Computational Infrastructure (NCI). These groups represent major generators of human genomic sequencing data in Australia and have varying levels of experience in submission to controlled access data archives. We identified key users and user stories for interaction with international repositories. The full set of user stories are found in Appendix B. In this analysis we focused on users who interact with an archiving platform and identified three main user types:

1. **Data Submitters**. Users responsible for assembling and transforming the metadata, encrypting human sequence data files and uploading data to an online repository.

2. **Data Consumers.** Users who intend to access and reanalyse a dataset for their own research use. These types of users will require access to either download or perform computation on a dataset, once appropriate approvals for access of data is obtained.

3. **Data Custodians:** Users responsible for ensuring storage, sharing, and use of the data occurs according to the consent granted by data donors, as well as legal, ethical, and social requirements. Data custodians may form part of a Data Access Committee (DAC) who will be required to provide long-term requests for data access. These users need long term support as any request for access to the data will need to be reviewed by the custodians and/or DAC.

In this section we discuss needs and use cases for each user type mentioned above. We refer to the users' experience with the EGA specifically as this is the main international archive that is used by Australian researchers and where the partners of this project have experience in submitting, downloading, and managing data access. It is expected that many of the same challenges would exist for other comparable international repositories.

Data Submitters

Data submitters have a number of requirements from a data repository service. In order to assemble a complete submission, they need to collect sample and sequence metadata, encrypt data files, and upload and register the dataset with the repository. Data submitters often submit to data repositories on behalf of others and can be under pressure to complete a submission within a tight deadline. Researchers usually require an accession given by archives, such as EGA or dbGaP, to prove their data has been submitted to an accessible repository before a manuscript can be submitted or published (e.g., Nature, 2022; PLOS ONE, 2019; Springer Nature, 2022). Therefore, the speed and ease of the submission process are important requirements for a data submitter. However, a successful dataset submission is not usually straightforward and can be quite onerous [30], particularly for those new to the process. Data submitters may range from highly technical users who want to automate programmatic methods of submission through to users requiring more user-friendly, web-based approaches. Therefore, data repositories generally need to support multiple methods of data and metadata submission.

The main challenges faced by Australian data submitters are:

- Assembling metadata and transforming into the required format.
- Encrypting and transferring data to the repository.

Metadata assembly, transformation and submission

Before data submitters can submit metadata to a data repository, they need to understand the structure and requirements of the metadata required for a submission. Newcomers to the process often need to invest considerable time and effort into first understanding these requirements, gathering the required metadata, then transforming the metadata they have into the required format, before submitting. Specifically, it can be difficult to understand the full list of metadata options available, and the provided controlled vocabulary options do not always accurately capture what a submitter wants to communicate about their dataset. For example, when submitting to the EGA, submitters can choose a 'Study Type' that describes their study, and although there are 14 descriptive options, 63 out of 68 studies submitted by Australian researchers have 'Other' as their study type.

Once data submitters understand the requirements, acquiring the metadata for the submission is the next challenge that they will face. As submission to a repository is generally thought of at the end of a project, collecting and storing the required metadata for a submission at the time of collection is not frequently done in a structured or routine manner. This creates a burden on the submitter to chase up the required information from a number of sources, such as biobanks, sequencing centres, research scientists, and laboratory assistants. Greater awareness of required information and incorporation into data collection protocols and management plans from the outset could help ease the metadata assembly process for the data submitter.

After the required metadata has been gathered, it needs to be transformed into the format required by the data repository. Depending on which database they want to submit to, this may be in the form of XML or JSON documents, tabular formats (e.g., excel spreadsheets or CSV files), or through manual web-based data entry [24,52–54]. To aid submission, institutes and individuals who regularly submit to data repositories usually create custom tools for programmatically creating and submitting this metadata. This is necessary because each submitter has a unique method for storing the required information, meaning tools must be tailored to their infrastructure. Having unified tools that could be used by existing, regular submitters would ease the burden of maintaining many tools and help newcomers who do not have resources to develop their own tools. Within Australia, this could be facilitated by large-scale adoption of standards-based solutions for metadata storage (e.g., FHIR [55], OMOP [56], and Beacon [57]), upon which unified tools can be built. An Australian-based archiving solution could also be customised to any Australian-adopted standards, facilitating data and metadata archiving submissions.

In the case of the EGA, there are two main methods of submission: via their interactive web-based submission portal [58] or programmatically via their REST API [59]. Submission via the portal is the route recommended by the EGA for most submissions. In this portal, the data submitter can either manually enter the information via a web-based form or fill out and upload a CSV template for each type of required information. Manual entry using the form works well when there are a limited number of entries required, however the burden increases with the number of objects that need to be submitted. This can be tedious, time consuming, and error prone for large numbers of sample and file objects.

Submitting using CSV templates requires the data submitter to transform the relevant metadata from their own structure into the required CSV columns. This can be challenging because there is little guidance on what the columns mean or how they should be populated. An important example is that it is not obvious that the ‘alias’ field for each object is an important field used for linking between the objects and needs to be globally unique within the submission account.

Programmatic submission via the REST API is an option for technical users and has the benefit of being able to be automated. However, successful submission using this method requires an in-depth knowledge of the XML schemas[60] defining the EGA metadata model and how to generate valid XML documents for each object of the submission[59]. This creates a barrier to understanding the full list of metadata options available and can lead to metadata not being populated accurately. In addition, it is difficult to understand how objects need to be linked. This makes the process of accurately linking metadata objects, such as samples, with their derived data files challenging.

Project partners QIMRB and CCIA have written custom tools that transform metadata from their systems to XML and submit them to the EGA via the API. This required considerable effort to set up initially, but has enabled a streamlined process for subsequent submissions. One pain point identified by the CCI team is that metadata objects submitted via this method are not visible within the main submission portal. Visibility of the objects within the portal would enable a data submitter to show the uploaded objects to their collaborators in a user-friendly way for validation before submission.

Other general use tools have been developed to assist with the process of EGA metadata creation. The EGA have developed a tool called ‘star2xml’, which converts from a multi-tabbed excel spreadsheet into the XML documents required for a submission [61]. ‘EGAsubmitter’ is a package that can guide a submitter through the whole process from metadata creation and submission to file encryption and transfer EGA [31]. It uses a combination of YAML and CSV files that it transforms to JSON for submission through the EGA JSON REST API [31].

Data encryption and transfer

Before sensitive human data can be uploaded to a managed access data repository, it must be encrypted [53,62,63]. This requires considerable computational resources to perform efficiently, particularly for large data collections. There is also a need for double the amount of storage space while the process is occurring (for both the encrypted and unencrypted versions of each file).

To submit to the EGA, data files must be encrypted using EGACryptor [62] before being uploaded to a user's 'submission box'. One pain point identified by the project is that all files need to have been encrypted before being uploaded as a batch. One way to mitigate this would be the ability to upload each file as it completes the encryption process. This would remove the need to have double the total storage capacity for a dataset available during the encryption process. This may be resolved with the release of a soon-to-be updated submission portal that incorporates 'on-the-fly' encryption [64].

Data files can be uploaded to the EGA via FTP or Aspera[65]. Due to the distance between Australia and Europe, this is prone to slow speeds and connection issues. QIMRB also regularly spends multiple days or weeks uploading a standard dataset. It is not unusual for the file transfer connection to drop out several times during a transfer process.

When preparing a submission to the EGA, all data files must be uploaded to a submission box. The EGA imposes a limit of 8TB for these submission boxes. This presents challenges for data submitters when the data they want to upload to a single study exceeds this size. It then becomes a slow, iterative process of uploading batches, waiting for the EGA to shift data before more data can be uploaded to the submission box. This exacerbates the already brittle upload process and depends on manual actions by helpdesk staff. As staff are located in a different time zone and support many users, the turnaround time for each iteration can be unpredictable, potentially spanning weeks to months. For example, an upload by CCIA of 300 TB took approximately 10 months to complete in 2020. Gaining a better understanding of data uploaders' needs and space requirements is an important consideration for a national solution to human genomic data archiving. Provision of flexible quotas could be one option to help prevent submitters from being hindered by reaching maximum quotas.

Globally, there is a trend toward tighter data privacy and security laws that protect where data can be stored[66]. In particular, data derived from clinical genomics tests, which are predicted to outstrip research-generated data in the near future[1], often have stringent jurisdictional legal and ethical requirements [19,67–69]. Currently there is no solution for long-term storage and sharing of human genomics data in Australia. Therefore, Australian data submitters that have the need for their data to remain within Australia, do not currently have an option for human genomic data archiving.

Data Consumer

Data consumers seek suitable datasets based on criteria of interest relevant to their research. Once a dataset of interest is identified, there is an approvals process, where the data consumer must request access from the Data Access Committee (DAC) and agree to the terms of data use before being granted authorisation to access the files. If approval is given, consumers' priority is to access and analyse the dataset in place, or download the dataset of interest as quickly, reliably and efficiently as possible.

The main challenges that Australian data consumers face are:

- Finding and accessing data that meets their criteria of interest.
- Accessing and analysing the data.

Finding and accessing data

Data consumers may use multiple methods to discover data that is held in repositories. One common way that data consumers find data is associated with published, peer-reviewed journal articles that cite an accession. As the data has been described by a publication, it is likely that there is enough detail to understand whether or not the dataset is of relevance to them. The data consumer can then visit the repository, find the dataset of interest, and begin the application process.

Alternatively, a data consumer may go directly to repositories that hold data of the type they are interested in and perform a search based on criteria of interest, such as disease, phenotype, assay type or study design. The data consumer may need to visit multiple repositories (e.g., EGA, dbGaP, JGA, GSA-Human) to find data that meets their requirements. Each repository uses different data models and terminology for the study-level metadata. This can make it challenging to find datasets that meet a data consumer's criteria of interest efficiently across these different data repositories. Consumers' criteria of interest are generally aimed at a research question, and therefore may require detailed clinical, phenotypic, or demographic information to determine suitability of the dataset. If this information is not provided by each repository in a structured format, it is difficult to narrow down relevant studies without a manual process of reading free text study descriptions or linked publications.

For example, the sample metadata submitted to the EGA requires biological sex and an uncontrolled description of 'phenotype' [58]. The description of phenotype is filled in quite inconsistently by Australian researchers (Figure 8) limiting the ability of data consumers to use this field to discover datasets of interest. Therefore, the primary way to search the EGA data portal is through keyword searching and relies on how the data submitter described their dataset in free text language.

Due to these minimal metadata standards, data consumers often need to base their decision on whether to go through the tedious data access process based on limited information. A more powerful query engine would allow data consumers to gain a more comprehensive understanding of the type of data and metadata available in a dataset and expedite the data discovery process. It would also help streamline the data access methods employed by a consumer, preventing the potential wasted effort of going through the process only to find the data is unsuitable. Improved searching through ontology expansion or the use of improved Artificial Intelligence (AI) language models are two options that could be utilised to improve data discovery within data repositories.

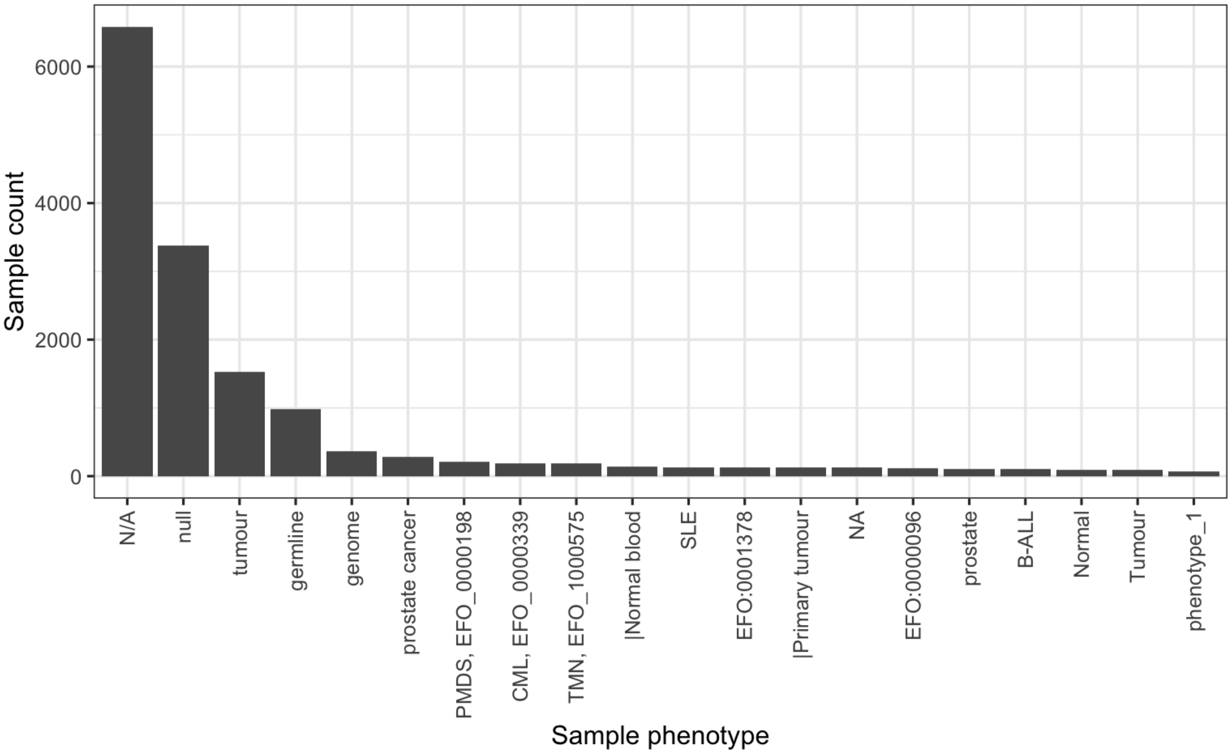


Figure 8. ***Top 20 phenotype terms used to describe samples in the EGA for Australian submitted data, retrieved via the EGA metadata API (2022-09-09).***

If access to the data is granted by the DAC, gaining access to detailed contextual metadata can also be challenging. It is possible to upload additional phenotypic files as part of an EGA submission, but no standardised format is prescribed. This can make it difficult for a data consumer to understand and integrate this information with their own data or other downloaded datasets. It is also common for no additional metadata to be uploaded alongside data files. This places a burden on data consumers to chase up this metadata directly from data custodians and/or from publications. Adoption of the 'Phenopackets' [70] format is a potential solution, providing a standard way of describing disease and phenotypic information, and can be currently uploaded as an 'Analysis' file to the EGA.

Accessing and analysing data

One of the biggest requirements from data consumers is an easy, reliable and fast way to access data from international repositories to reanalyse for their own research. A major impediment to this is the distance between the location of the data and Australia, which impacts download speeds and reliability. The current system is prone to network issues, slow speeds, and regular cut offs. This makes it difficult for Australian researchers to access both Australian and International datasets stored in international repositories, such as the EGA and dbGaP.

Data consumers also have a need to access only the data that is relevant to their analysis to avoid wasted effort, storage space, and time. In the EGA, access is granted per dataset, and although it is possible to download specific files and even particular genomic ranges from files in a dataset [71], the limited available metadata makes it challenging to determine which files would meet specified criteria before downloading. More detailed and accurate metadata could help streamline this process and allow users to only download or access the files and file segments of interest.

Enabling pipelines to be run directly on the data could be another approach to meeting consumers' needs for a more efficient and streamlined analysis process. This would remove the need to download large data files and instead allow for smaller results files to be downloaded.

Data Custodians

Data custodians are often part of the DAC and they require long-term support to grant access to the dataset each time a data access request is made. Their needs revolve around being promptly notified about incoming requests, assessing whether the application meets the Data Access Policy for a dataset, and if successful, having a convenient way to approve these requests. Although the notification and granting of access is usually managed within the data repository's system, data custodians will have unique data access policies and processes that must be followed before access is granted. The policies and processes tend to vary by institute, consortium, or dataset. They may also use external DAC management systems such as Resource Entitlement Management System (REMS) or Data Use Oversight System (DUOS) [72–74].

One main requirement for this user type is the ability to easily grant access to data under their custodianship with collaborators after uploading to a data repository. This will help to encourage a more cohesive way to share data that is within the specified data access and use policy.

In the case of the EGA, Australian data custodians and data consumers will often find alternate means of sharing outside the EGA system, even when the data has been successfully archived and found via the EGA portal. This is due to the long process involved in granting access, as well as the difficulties of Australian researchers needing to download data from the other side of the world. It is likely that if the process to request and grant access was easier and the means to access the data was more reliable, researchers would use a data repository's access system, rather than finding workarounds where legal, ethical, and duplication issues may arise.

Another need for data custodians is the ability to share and provide access to specific subsets from broader datasets specified within data repositories. Within the EGA, a set of samples and their data is batched together and belongs to one dataset. Having the ability to grant access to a subset of their main dataset may be in the data custodians' best interest and may be an ethics requirement in some cases.

Data custodians would benefit from tools to help manage the data access request process. Currently, many custodians rely on a manual, email-based system that is difficult to track and audit. Software tools that facilitate consistent annotation and systematic categorisation of both datasets and users would enable a more streamlined process. It could also allow access decisions to be supported by automated recommendations. A more in-depth understanding of candidate solutions for software that aid the Data Access Request process is being gathered in the HGPP's DAC Automation sub-project. That sub-project Discovery Phase Report is published on Zenodo[3].

Recommendations arising from the Australian Community needs analysis

An Australian FEGA node could help meet the needs of data submitters through:

- Dedicated Australian-based helpdesk for more timely and responsive support, documentation, and targeted outreach and training.
- Potential for customised policies and processes that suit the Australian research landscape.
- Potential for unified-tools development to facilitate submission based on nationally-adopted standards.
- Improved reliability and speed for data uploads and downloads.
- Data that cannot leave Australia can be archived on-shore.
- Potential flexibility and customisation of quotas for data upload areas.

An Australian FEGA node could help meet the needs of data consumers through:

- Improved data discovery through support and encouragement to meet better metadata standards, supplemental metadata model, and improved portal.
- Improved reliability and speed for data downloads of Australian data.
- Alternate, flexible data access (e.g., streaming, cloud compute integration).
- Potential mirroring of high-value international datasets that are frequently accessed by Australian researchers.

An Australian FEGA node could help meet the needs of data custodians through:

- Access to dedicated Australian-based helpdesk for improved responsiveness.
- Establishment of a software assisted DAC system (e.g., REMS to manage data access requests).
- Ability to grant access to data subsets.

[3] https://zenodo.org/records/6644050

Legal considerations

There are many legal considerations that need to be explored before setting up a national data repository of this type within Australia. Consultations with existing European federated nodes have emphasised that this process can take considerable time due to complex legal arrangements[75,76]. Expert legal advice specific to the Australian landscape needs to be sought in order to ensure compliance. Some specific considerations that need to be further explored are summarised below.

1. Whether the data being held within the archive is considered ‘personal information’ and governed by the Privacy Act[77].

   Although Human Genomic data itself may not be considered ‘personal information’ and covered under Privacy Act[19], the combination of this data alongside phenotype, clinical, and demographic information could be considered ‘personal information’ if it is used to re-identify an individual.

2. How to ensure the legal responsibility of the data remains with the data custodian if the data is being held by the repository.

   In Europe’s GDPR legislation[78], the role of ‘data processors’ and ‘data controllers’ are clearly defined. This provides the ability to create ‘data processing agreements’ between data controllers and data processors. In the case of the EGA, this ensures that data custodians maintain control of the data and are considered the ‘data controllers’ as they make the decision on who can access the data. Whereas the EGA team takes on the role of ‘data processors’ in order to store and make the data available based on decisions made by the data custodians.

   In Australia, the Privacy Act 1988[18,77] does not have these concepts outlined, so it is unclear how a similar legal arrangement would be reached between submitters and the national archive.

3. What legal entity will be the host for the national archive?

   The repository will need to be able to sign contracts to enable data processing and data distribution. It would need to either be itself a legal entity, or be housed within a body that is willing to act as its legal entity.

3. Recommendations on Preferred Approaches and Technologies

Based on the currently available options to establish a national data repository for long-term storage and sharing of sensitive human ‘omics data, we recommend the use of the FEGA model as a core component of a controlled-access national ‘omics data repository.

This model offers several advantages, including:

- It builds upon the Central EGA’s vast experience over 15 years in running repositories of this type.
- There is an active, global, and growing community of support.
- It underlies national infrastructure in several countries, so it is likely to continue to be maintained.
- There is existing, well-maintained software to build upon and customise.
- It is fully deployable locally with full control over hosting of the platform and the data.
- There are existing guidelines, policies, standards, templates, and advice to access and build upon.
- It provides an internationally-recognised accession accepted by publishers.
- It is a familiar platform, already in use by many human genomics researchers in Australia.
- It will allow more efficient data access and submission for Australian researchers.
- It would facilitate broader discovery of Australian research data through connection to the central EGA (versus setting up a stand-alone Australian repository).
- It provides a starting point and solid foundation to build future improvements and to integrate with future national scale human ‘omics infrastructure.

There are several options for the implementation of a national ‘omics data repository, which includes a FEGA node. Below we discuss potential implementation options and high-level resourcing requirements.

FEGA Evaluation

Technical implementation

At the time of writing this report, there were two different implementations of the FEGA local software stack:

1. LocalEGA (provided by Central EGA as a reference implementation).
2. NEIC Sensitive Data Archive (SDA, a full implementation as used by NEIC partner sites).

The main components of a FEGA local software stack include:

1. Inbox for receiving (encrypted) files from researchers.
2. Archive storage for storing files long term.
3. Backup storage for storing an independent duplicate copy of files long term.
4. Processing logic to verify file integrity and move copy to archive and backup storage.
5. A message queue (bi-directional) to communicate with Central EGA.
6. A data access tool to allow approved researchers access to selected datasets.

To evaluate each implementation, we installed each on a VM hosted in the NCI Nirin Cloud. We ran the self-testing scripts, attempted to ingest some sample data, reviewed the code-base and the activeness of the development community around it.

LocalEGA Reference implementation:

- Installation: Was packaged in a Docker container. This meant that all dependencies and necessary functionalities were packaged together in the container to help with software installation. We were able to successfully run the installation with no reported issues.
- Testing: All self-test scripts returned a failure when run after the installation. Investigating this further would have taken a considerable amount of time to debug and apply code changes.
- Data Ingest: As a proof of concept, we were able to successfully ingest a simple text file.
- Code review: While the implementation was quite simple, it did include a fork of SSHD in the SFTP inbox component, which was taken from 2019 and not actively developed. This represents a significant security risk and needs to be addressed for a production implementation. Given that the upstream repository does not maintain it, the local deployment organisation would have to factor in extra development time to maintain it or implement their own inbox component.
- Developer community: Appears to be mostly a single developer and has provided little development since mid-2020.
- Reference: https://github.com/EGA-archive/LocalEGA [79].

NEIC SDA:

- Installation: Relatively easy to install an instance because it comes with a docker-compose file that deploys reference implementations of all required components. There was a small issue getting the self-signed certificates to all function correctly under firefox, however these would be replaced with certificates signed by a public-trusted certificate authority on a production system. The interfaces that require the certificates are only internally used.
- Testing: Local testing functioned as expected, including a sample script that ingested a sample file through all steps of ingestion.
- Data Ingest: We were able to simulate our own ingestion and simulate tasks that would be performed through the Central EGA portal.

- Code review: No major issues were identified with the code review; no unsupported third party libraries and an active developer community with regular code updates from multiple developers. The code-base supports multiple storage platforms (object or block storage).
- Developer community: Is very active and consists of a number of engaged developers, as would be expected from an implementation used by multiple countries in their Federated EGA node.
- A lot more active development in this code repository in comparison to the localEGA reference implementation.
- Other noteworthy features:
  - Uses REMS as the preferred tool for DACs, which is recommended by other HGPP sub-projects.
  - Designed so that it can function as a stand-alone sensitive data archive and be a single system that handles both EGA and non-EGA data.
- Reference: https://github.com/neicnordic/sda-pipeline[80].

Overall, the NEIC SDA is the recommended option since it has an active developer community and is used in production in Sweden, Norway, and Finland.

Minimal Viable Product (MVP) - Australian FEGA node

Here we define a minimal viable product that we believe would meet a large number of requirements of the Australian human 'omics research community. While it may not fulfil all community needs identified in this report, we consider it a significant step forward to enhance Australian researchers' ability to securely archive their human 'omics data and metadata, and to enable greater discoverability and accessibility of Australian research data. The MVP should aim to reach recommended levels of the FEGA maturity model as a starting point [37,38] (see Appendix C).

The host of an Australian FEGA node (AUSFEGA) would need to have considerable skills and experience in the provision of national-scale digital infrastructure. Security assurance of the platform will be paramount to ensure that all consent, ethical, and legal considerations are adequately met. We would also recommend that AUSFEGA utilise standard interfaces ensuring it is well-connected to existing national digital infrastructure and cloud platforms to facilitate efficient data movement and analysis. We recommend the node uses federated identity management through, for example, the AAF and CILogon.

This MVP solution (Figure 9) would continue to rely heavily on the services provided by the Central EGA, including their submission, discovery, and DAC portals. Subsequently, this also means a reliance on the services of the Central EGA's helpdesk. However, this offers the advantage that fewer resources would be needed to get an AUSFEGA initiated. Resources within the AUSFEGA would be able to focus on deploying the systems and support required to store human genomics data in Australia and liaise with the Central EGA.

Although we refer to a single AUSFEGA node, it is possible that multiple data storage locations exist in a federated solution within Australia. This would be feasible provided all nodes follow the principles set out by AUSFEGA and adhere to standards that allow seamless integration between storage nodes and national and international computational infrastructure.

Given the Central EGA is provided as a free service, we would recommend the AUSFEGA node be fully funded and provided at no cost to Australian researchers for submission or access. This requires significant long-term investment. Other FEGA nodes have been established based on large national government grants, such as the GHGA [81] and often supported by European funding (e.g., ELIXIR and Horizon 2020 [43]).

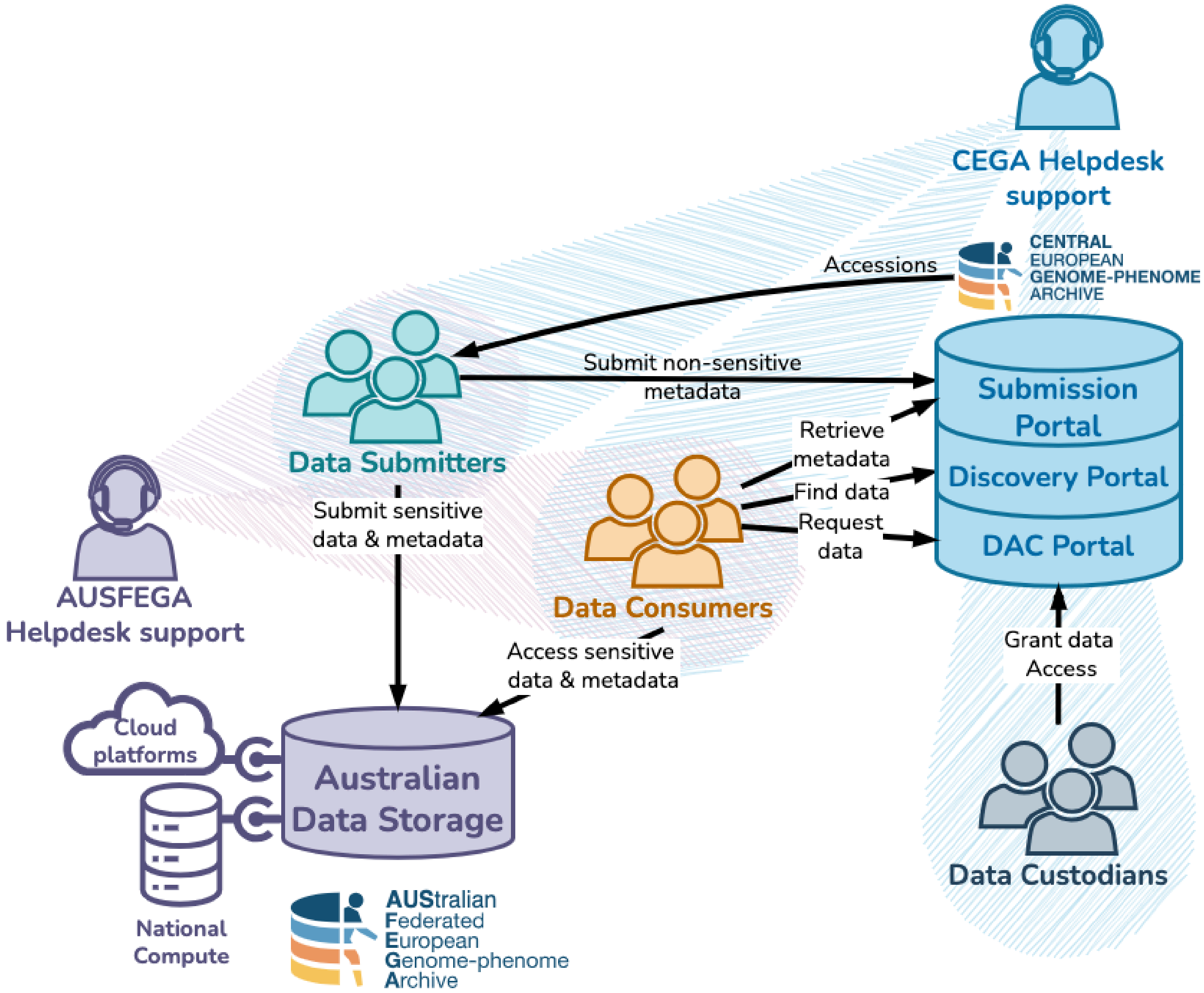


Figure 9. ***Structure of an MVP Australian Federated EGA node (AUSFEGA). In this MVP, setup users continue to rely on the Central EGA for submissions, discovery, and data access management, and receive support from both Central EGA and AUSFEGA helpdesk staff. All sensitive data is stored in Australia and connected to cloud platforms and digital research infrastructure. Data storage may be centralised or held in multiple trusted locations.***

Main use cases addressed with the MVP Solution

- Improved speed and reliability of data transfer when submitting or accessing data for Australian researchers.
- Internationally recognised accessions through the EGA that:
    - Facilitate publication in peer-reviewed journals.
    - Fulfil funding requirements to deposit data in recognised archives.
- Long-term storage in Australia (minimum four-year commitment as FEGA node[82]).
- Improved data discoverability for Australian datasets, nationally and internationally.
- Better support, awareness, training, and outreach that caters directly to Australian researchers within their time zones.
- Ability to share data that must remain within Australia.

Infrastructure requirements

As a federated node of the EGA, the AUSFEGA node would need to incorporate an implementation of the LocalEGA software. This software communicates with the Central EGA in order to assign and send globally-unique accessions to sensitive data files that are housed within the federated EGA nodes' storage systems. At the time of writing , the NeIC SDA[83] was the most mature implementation, however this should be revisited at the time of implementation to ensure it remains the leading implementation. This implementation provides the infrastructure for submission, ingestion, and data retrieval of files stored within the federated node. Submission of non-sensitive metadata would occur through the Central EGA submission processes.

The data access model would be similar to the current Central EGA where users would gain access to datasets through a data access request being assessed by a DAC. Data access requests would be managed through the Central EGA DAC portal. Data access would be on the level of a dataset with connection to national computing and cloud services allowing efficient access for Australian researchers to stream or download in comparison to downloading from overseas.

The compute resources needed to run the software required to support a FEGA node includes[36,76,84]:

- Compute resources
    - Running the LocalEGA software
    - Integrated with Federated Identity management (e.g., CILogon)
    - Running operational services
        - Issue trackers for helpdesk, operations, development
        - Support, ETL, encryption/decryption, QA
        - Dev/testing environments
    - Documentation and SOPs
- Network/transfer
    - Means to efficiently transfer data to Australian research institutes, national compute, cloud (e.g., Aspera, FTP, Globus)

- Storage resources
    - Staging storage
        - Upload space for submitters with unique submission boxes
    - Permanent archival storage
        - Ideally with two geographically-distinct copies
        - Object-based, allow cloud/streaming access

Staffing requirements

Governance and Legal

Within the GDPR legal framework[85], there is a need to have a Data Protection Officer (DPO), which is specified within the FEGA maturity model[37; 1.1.1, Level 3] and FEGA Node operations guidelines[84]. While DPOs are not specified within the Australian legal system[85], it is likely that a similar position will be needed to advise and coordinate with other legal experts to establish data processing procedures, contracts, and legal agreements between the data custodians and the organisation running the archive.

Resource documentation will be needed that includes patient consenting frameworks compatible with the data sharing agreements that the archive will establish. This is required to support patient cohorts that can be shared because the cohort policy and patient consent is appropriate, and is mainly relevant for researchers/clinicians recruiting new cohorts. Data custodians with existing cohorts will need to check if their existing consent frameworks are compatible with data sharing before submission. We do not recommend allowing for storage of data that cannot be shared.

Development

Freeberg et al.[36] estimates four to six people (or two to three Full Time Equivalent (FTE)) will be required to resource the development needed to establish an FEGA node. The tasks for this team would include:

- Deploying existing software on Australian infrastructure.
- Establish data storage and access processes.
- Minor customisation.

Based on the required FTE estimations and an average 1.0 FTE rate of AUD $160K (including overhead costs), setting up the system would cost approximately AUD $320K-$480K in staffing.

Operations

Resourcing of four to five FTE spread across six to eight employees is recommended for operating and maintaining the archive software after it is established[36]. The spread across multiple employees ensures there is an adequate breadth of domain expertise and background to cover the diverse needs of the platform. We recommend at least two FTE in the operations team be focused on helpdesk activities to support data submitters with onboarding their cohorts into the archive. Ongoing staffing costs are estimated to be AUD $640K-$800K per annum.

Community Engagement

Community engagement and training is essential in order to build awareness, uptake, and use of the genomics data archive. Given the EGA is already a preferred archive location for Australian researchers (discussed earlier in this report), we anticipate that the Australian genomics community will enthusiastically adopt an Australian-based genomics archive. We would recommend up to one FTE (AUD $160K) be focused on community engagement and awareness building through activities such as:

- Gathering needs, requirements and feedback.
- Workshops and training.

MVP+: Integration with Beacon v2 and REMS

The MVP+ option would build on an established MVP. On top of an MVP, integration with further services would improve the utility and usability of an AUSFEGA node (Figure 10).

Beacon v2 is a GA4GH standard that provides a common data model and ability to search for data based on demographic, clinical, and phenotypic information, as well as genomic variants [57]. FEGA-Beacon v2 integration would help with discoverability of datasets through improved querying and metadata quality. This would give researchers a clearer idea of the available data and assess its relevance to them before embarking on the data access process. If they do obtain access, the enhanced metadata would help researchers pinpoint the specific files they need, making their analyses more efficient. The data and metadata within Beacon may even provide enough information for a researcher to avoid the need to gain access to the full raw data, significantly reducing their analysis and storage costs. A FEGA-Beacon v2 could also allow integration into broader national and international Beacon networks, further enhancing discoverability.

Data submitters may need additional support to provide the enhanced metadata required for Beacon v2, as well as tools to facilitate submission in the correct format. There would also be a need for development and maintenance of a web-based portal to enable a user-friendly method for querying the Beacon v2 API.

The data access request process would be facilitated with a REMS[74] instance to enable semi-automation and improved auditing. REMS is a web-based DAC management platform that was explored as part of the DAC Automation sub-project within the HGPP. Further information about how it was tested and accepted as a viable technology for DAC management in Australia can be found in the DAC Automation Pilot Implementation report.

Integration with data support software would allow more flexible and customised access to data. This would allow custodians to decide to give access to specific files or genomic regions based on the specific needs of the data consumer. Providing access in this way could meet stricter legal, ethical, and consent requirements, such as dynamic consent, and adheres to the five safes principles by only providing access to the specific data that is requested. A program called ELSA[86] is currently being developed to allow this type of access and may provide a solution to more flexible and automated data releases from a national 'omics repository.

Additional use cases fulfilled

- Improved, streamlined, and customised DAC management.
- Improved metadata and variant searching.
- Improved discoverability and virtual cohort assembly.
- Integration of metadata and variants into national and international networks.
- Ability to share specific data subsets.
- Improved automation of data releases.

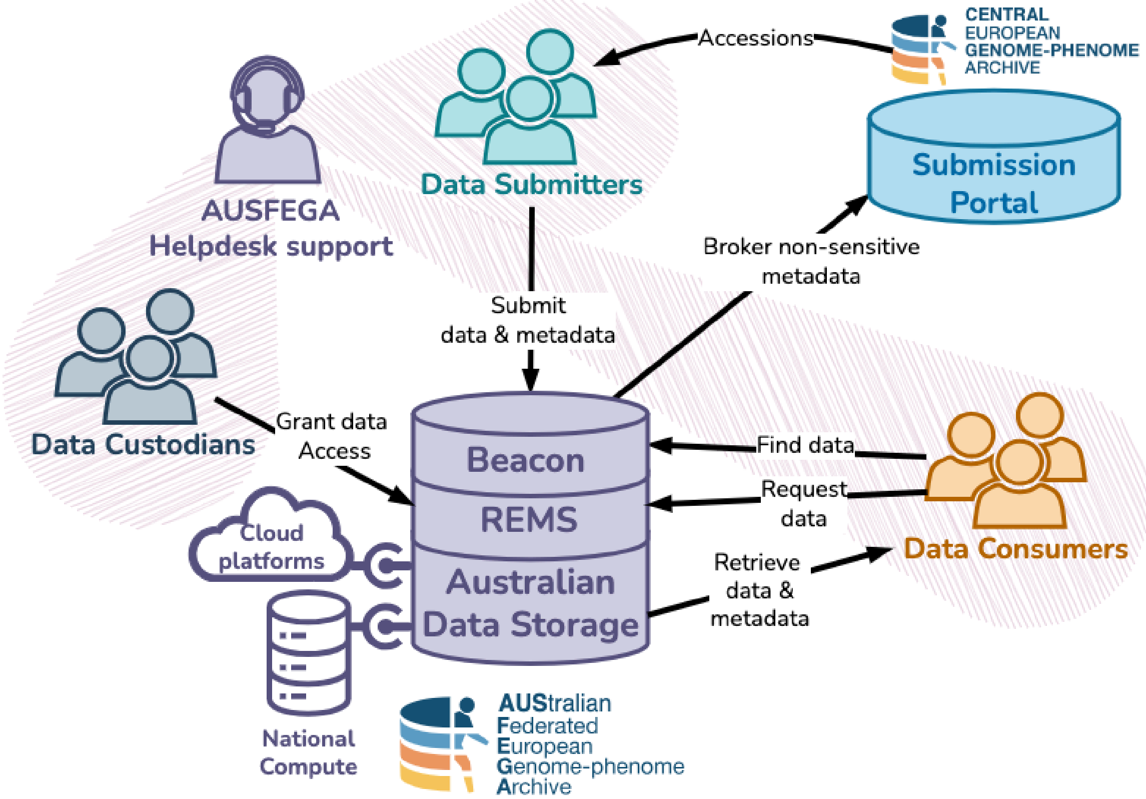


Figure 10. ***Structure of an MVP+ Australian Federated EGA node (AUSFEGA). In this enhanced setup, users interact with AUSFEGA for all tasks. Beacon provides enhanced discoverability and metadata, while REMS offers flexible data access management. AUSFEGA help desk provides support to all users. Required EGA metadata is brokered directly from AUSFEGA to Central EGA.***

Additional infrastructure requirements

- Compute resources

    - Running REMS instance for Data Access Management.
    - Running Beacon API implementation.
    - Running web portal for user-friendly querying.

Additional staffing requirements

- One FTE establishing and maintaining Beacon instance and web portal.
- One FTE establishing and maintaining REMS instance.
- One FTE help desk support for Beacon and REMS.

National Federated Omics Platform

A national-scale, controlled access human ‘omics repository would form one part of an overall national approach to human ‘omics data management. However, there are many other components needed in order to make it a functioning and useful ecosystem that progresses the overall goals of Australian human ‘omics research. The recommendations here seek to ensure that any national human ‘omics repository fits within any future national digital research infrastructure, as well as the broader international landscape, by advocating for the use of international standards and existing, well-adopted standards.

4. Conclusions and Future Directions

The amount of data generated from human genomics research in Australia continues to grow at a rapid pace. Australian researchers are facing difficulties in managing the archival storage of these data and associated metadata. Some platforms do exist to solve these problems, however they are not ideal for many Australian researchers. In addition, there is no unified national archive designed to support the Australian genomics community, nor is there a set of standardised guidelines on how to archive genomic data and metadata.

The data and metadata archiving sub-project has:

- Investigated the available options and requirements for human genomics data archiving and metadata.
- Gathered community needs from the representatives of the HGPP project partners.
- Provided recommendations on how to fulfil community needs.
- Defined an MVP, as well as additional features that can be added downstream as part of an MVP+ solution.

Based on the findings of this report and the available options at the time of writing, we recommend the use of the FEGA model as a core component of the Australian genomic data repository for archiving and storing human genomics data. The MVP solution would address many but not all user requirements, however, it is a giant leap forward and has the ability to address further requirements by integrating it with other HGPP components, such as the DAC and Virtual Cohorts sub-projects.

In addition, we have highlighted here some resource requirements for the establishment and operation of a FEGA node. Some key considerations are the infrastructure requirements, as well as the staffing requirements. Establishing one or more FEGA nodes in Australia will greatly benefit many Australian researchers. Setting up and maintaining each node will need significant ongoing resourcing. We are hopeful for future funding opportunities to support the establishment and maintenance of this infrastructure, which will in turn help advance human genomics research and discovery in Australia.

5. References

1. Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: GA4GH looks to 2022. 203554 Preprint at https://doi.org/10.1101/203554 (2017).
2. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
3. Corpas, M., Kovalevskaya, N. V., McMurray, A. & Nielsen, F. G. G. A FAIR guide for data providers to maximise sharing of human genomic data. *PLOS Comput. Biol.* **14**, e1005873 (2018).
4. FAIRsharing team. FAIRsharing | Home. https://fairsharing.org/ (2022).
5. van der Velde, K. J. *et al.* FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. *Sci. Data* **9**, 169 (2022).
6. Vesteghem, C. *et al.* Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. *Brief. Bioinform.* **21**, 936–945 (2020).
7. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
8. Nature. Data Repository Guidance | Scientific Data. https://www.nature.com/sdata/policies/repositories (2022).
9. PLOS ONE. Data Availability. *Data Availability | PLOS ONE* https://journals.plos.org/plosone/s/data-availability (2019).
10. Springer Nature. Mandated data types | Authors | Springer Nature. https://www.springernature.com/gp/authors/research-data-policy/repositories-socsci/19540364 (2022).
11. NIH. NIH Genomic Data Sharing Policy. https://sharing.nih.gov/genomic-data-sharing-policy (2023).
12. Wang, Z., Lachmann, A. & Ma'ayan, A. Mining data and metadata from the gene expression omnibus. *Biophys. Rev.* **11**, 103–110 (2019).

13. NHMRC. Australian Code for the Responsible Conduct of Research, 2018 | NHMRC. https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018 (2018).

14. NHMRC, ARC & Universities Australia. *Management of Data and Information in Research: A guide supporting the Australian Code for the Responsible Conduct of Research*. https://www.nhmrc.gov.au/sites/default/files/documents/attachments/Management-of-Data-and-Information-in-Research.pdf (2019).

15. Department of Health. *National Health Genomics Policy Framework 2018–2021*. https://www.health.gov.au/resources/publications/national-health-genomics-policy-framework-2018-2021?language=en (2017).

16. Queensland Genomics. *Blueprint for a National Approach to Genomic Information Management (NAGIM)*. https://queenslandgenomics.org/capability-initiatives/national-approach-to-genomics-information-management/ (2020).

17. Australian Genomics. *National Approach to Genomic Information Management (NAGIM) Implementation Recommendations*. https://www.australiangenomics.org.au/wp-content/uploads/2021/06/NAGIM-Implementation-Recommendations-December-2022.pdf (2022).

18. Belcher, A., Haas, M., Grewal, N. & Newson, A. Genomic Data & Privacy Law: A summary of Health Legal's report for Australian Genomics. https://www.australiangenomics.org.au/wp-content/uploads/2021/09/Summary-Health-Legal-Report.pdf (2018).

19. Paltiel, M., Taylor, M. & Newson, A. Protection of genomic data and the Australian Privacy Act: when are genomic data 'personal information'? *Int. Data Priv. Law* ipad002 (2023)

doi:10.1093/idpl/ipad002.

20. Freeberg, M. A. *et al.* The European Genome-phenome Archive in 2021. *Nucleic Acids Res.* **50**, D980–D987 (2022).
21. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975-9 (2014).
22. Okido, T. *et al.* DNA Data Bank of Japan (DDBJ) update report 2021. *Nucleic Acids Res.* **50**, D102–D105 (2022).
23. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. *Nucleic Acids Res.* **50**, D27–D38 (2022).
24. NIH. How to Submit A Non-NIH funded Study to dbGaP | Data Sharing. https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/how-to-submit-a-non-nih-funded-study-to-dbgap.
25. NIH OSP. Data Repositories and Trusted Partners. *Data Repositories and Trusted Partners* https://osp.od.nih.gov/policies/scientific-data-management-policy/data-repositories-and-trusted-partners/ (2023).
26. Kids First. phs002172.v1.p1 Gabriella Miller Kids First Pediatric Research Project in Microtia in Hispanic Populations. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002172.v1.p1 (2020).
27. Kids First DRC. Kids First Data Portal. https://portal.kidsfirstdrc.org/dashboard (2023).
28. Kids First DRC. Kids First DRC Help Center - Studies and Access. *Notion* https://d3b.notion.site/Studies-and-Access-a5d2f55a8b40461eac5bf32d9483e90f (2023).
29. Cliffe, A. Overview: Terra on Azure. *Terra Support* https://support.terra.bio/hc/en-us/articles/12028783864859-Overview-Terra-on-Azure (2023).
30. Band, G. Me vs. EGA.

https://gavinband.github.io/bioinformatics/data/2019/05/01/Me_versus_the_European_Genome_Phenome_Archive.html (2019).

31. Viviani, M. *et al.* EGAsubmitter: A software to automate submission of nucleic acid sequencing data to the European Genome-phenome Archive. *Front. Bioinforma.* **3**, (2023).
32. Pinese, M. *et al.* The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly. *Nat. Commun.* **11**, 435 (2020).
33. Freeberg, M. & Curwin, A. Federated EGA Updates in 2022. *F1000Research* **11**, (2022).
34. ISO. ISO/IEC 19464:2014 - Information technology — Advanced Message Queuing Protocol (AMQP) v1.0 specification. (2014).
35. Curwin, A. *et al.* Demonstrating federated EGA services for sensitive data discovery & access. (2023).
36. Freeberg, M. A. *et al.* Resource considerations for establishing and operating a federated human data node. *F1000Research* **11**, (2022).
37. FEGA. Federated EGA Maturity Model. https://inab.github.io/fega-mm/ (2022).
38. Capella Guitierrez, S. FEGA Maturity model task force Practical Implementations of FEGA Maturity Model. (2022).
39. Keane, T. 2022 Resolving your Federated EGA Legal Challenges. (2022).
40. NBIS. FEGA Sweden. https://fega.nbis.se/ (2023).
41. NBIS. SDA-CLI. https://github.com/NBISweden/sda-cli (2023).
42. SciLifeLab. NBIS signs the Federated EGA Collaboration Agreement. *SciLifeLab* https://www.scilifelab.se/news/nbis-signs-the-federated-ega-collaboration-agreement/ (2022).
43. CSC. A step towards safe access to sensitive human data across borders - CSC has signed the Federated EGA contract - CSC Company Site. https://www.csc.fi/en/-/csc-has-signed-the-federated-ega-contract (2022).
44. CSC. Finnish Federated EGA Node - Docs CSC. https://docs.csc.fi/data/sensitive-data/federatedega/

(2022).

45. Horn, R. & Merchant, J. Managing expectations, rights, and duties in large-scale genomics initiatives: a European comparison. *Eur. J. Hum. Genet.* **31**, 142–147 (2023).
46. GHGA. ghga-de/ghga-metadata-schema: Metadata schema for the German Human Genome-Phenome Archive (GHGA). https://github.com/ghga-de/ghga-metadata-schema (2022).
47. Hornos Vidal, A. *et al.* The Spanish Node of Federated EGA. *F1000Research* **11**, (2022).
48. NFEGA. Federated EGA Norway node. https://ega.elixir.no/ (2023).
49. NFEGA. *Expanded NFEGA service description*. https://ega.elixir.no/docs/Expanded%20FEGA%20Norway%20service%20description.pdf (2023).
50. NFEGA. *System description Federated EGA Norway*. https://ega.elixir.no/docs/System_description_FEGA-Norway.pdf (2023).
51. elixir-oslo. lega-commander. https://github.com/elixir-oslo/lega-commander (2023).
52. DDBJ. JGA submission steps. https://www.ddbj.nig.ac.jp/jga/submission-step-e.html.
53. Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T. & Ogasawara, O. DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res.* **49**, D71–D75 (2021).
54. NCBI. Submitting data - GEO - NCBI. https://www.ncbi.nlm.nih.gov/geo/info/submission.html (2023).
55. Alterovitz, G. *et al.* FHIR Genomics: enabling standardization for precision medicine use cases. *Npj Genomic Med.* **5**, 1–4 (2020).
56. OHDSI. OMOP Common Data Model – OHDSI. https://www.ohdsi.org/data-standardization/the-common-data-model/ (2022).
57. Rambla, J. *et al.* Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond. *Hum. Mutat.* **43**, 791–799 (2022).
58. EGA. Submitter Portal - EGA European Genome-Phenome Archive. https://ega-archive.org/submission/tools/submitter-portal (2023).

59. EGA. Programmatic submissions (XML based) - EGA European Genome-Phenome Archive. https://ega-archive.org/submission/sequence/programmatic_submissions (2023).

60. EGA. Prepare XMLs - EGA European Genome-Phenome Archive. https://ega-archive.org/submission/sequence/programmatic_submissions/prepare_xml (2022).

61. Casado Barbero, M. star2xml. https://github.com/EGA-archive/star2xml/ (2021).

62. EGA. EGACryptor. https://ega-archive.org/submission/tools/egacryptor (2023).

63. NIH NLM. dbGaP Study Submission Guide. https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/ (2023).

64. EGA. EGA submitter portal tutorial | VEIS - Valorización de EGA para la industria y la sociedad. https://veis.bsc.es/events/ega-submitter-portal-tutorial/ (2022).

65. IBM. Aspera. https://www.ibm.com/products/aspera (2023).

66. OAIC. Chapter 8: APP 8 Cross-border disclosure of personal information. *OAIC* https://www.oaic.gov.au/privacy/australian-privacy-principles/australian-privacy-principles-guidelines/chapter-8-app-8-cross-border-disclosure-of-personal-information (2019).

67. Health Legal. Genomic Data & Privacy Law: A summary of Health Legal’s report for Australian Genomics. https://www.australiangenomics.org.au/wp-content/uploads/2021/09/Summary-Health-Legal-Report.pdf (2018).

68. Lalova-Spinks, T., Meszaros, J. & Huys, I. The application of data altruism in clinical research through empirical and legal analysis lenses. *Front. Med.* **10**, (2023).

69. Warren, V., Critchley, C., McWhirter, R., Walshe, J. & Nicol, D. Context matters in genomic data sharing: a qualitative investigation into responses from the Australian public. *BMC Med. Genomics* **15**, 275 (2023).

70. Jacobsen, J. O. B. *et al.* The GA4GH Phenopacket schema defines a computable representation of

clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).

71. EGA. EGA download client: pyEGA3. https://github.com/EGA-archive/ega-download-client (2022).
72. Broad. DUOS UI. https://github.com/DataBiosphere/duos-ui (2023).
73. Cabili, M. N. *et al.* Empirical validation of an automated approach to data use oversight. *Cell Genomics* **1**, 100031 (2021).
74. CSC. Resource Entitlement Management System. https://github.com/CSCfi/rems (2023).
75. Englund, M., Viklund, J. & Nigru, S. Meeting with Finnish and Swedish FEGA nodes. (2022).
76. *Federated EGA webinar*. (2022).
77. AG. Privacy Act 1988. http://www.legislation.gov.au/Details/C2022C00199 (2022).
78. European Council. Chapter 4 – Controller and processor Regulation (EU) 2016/679 (GDPR). *General Data Protection Regulation (GDPR)* https://gdpr-info.eu/chapter-4/ (2016).
79. Local EGA main repository. https://github.com/EGA-archive/LocalEGA (2022).
80. sda-pipeline. https://github.com/neicnordic/sda-pipeline (2022).
81. GHGA. Letter of Intent within the Call for National Research Data Infrastructures (NFDI), S. 6. (2019).
82. Keane, T. *et al. EGA Federation: Structure and organisation v1.1*. https://ega-archive.org/files/EGA-Federation-Structure-v1.1.pdf (2020).
83. NeIC. NeIC SDA Operations handbook. https://neic-sda.readthedocs.io/en/latest/ (2023).
84. Kerry, G. *et al. Federated EGA Node Operations Guidelines*. https://ega-archive.org/files/EGA-Node-Operations-v2.pdf (2022).
85. Potter, A. *et al. Comparing privacy laws: GDPR v. Australian Privacy Act*. https://www.dataguidance.com/sites/default/files/gdpr_v_australia.pdf (2020).
86. Patterson, A. Elsa Data. https://github.com/elsa-data/elsa-data (2023).

6. Appendix

A. Current workflow to Archives - Case Studies

QIMRB

The QIMRB have a semi-automated process for preparing data and metadata for submission to the EGA. A step-by-step explanation of the process is described in here.

Firstly, files are manually inspected and edited to ensure they do not contain any potentially identifying information, with sample identifiers usually changed to match identifiers used in the publication. The EGA web-based submission interface is then used to create the study, DAC, and policy objects. The data policy links to a pdf document stored in figshare (example).

Custom scripts are used to encrypt data files using HPC resources. Encrypted files are then uploaded to the EGA staging area via Aspera or ftp. The overall size of the files can cause challenges as EGA submission boxes are limited to 10TB, meaning files need to be batched in order to complete upload of the entire submission. This can delay the submission process significantly. Between each batch, the files need to be fully archived before the next batch can be uploaded. This archiving process can vary in time as it requires manual intervention from Central EGA helpdesk staff, who typically have a large backlog of requests.

A custom tool is used to extract information from the bam or FASTQ files to prepare the analysis XMLs that are required for each file. Another tool combines information from the analysis XMLs, along with information from QIMRB's sample database, to create the sample XMLs. Sample information is double-checked with researchers before submission.

Submission of analysis and sample XMLs cannot proceed until all files have been uploaded. A tool is used to upload the XML files then update QIMRB's database with the returned EGA accessions. Data files and samples are grouped together within a dataset and submitted as an additional XML using custom tools. Once the entire Study is ready for public release, an email is sent to the EGA helpdesk to request it be released. This can take days to weeks depending on helpdesk availability.

UMCCR

UMCCR do not currently have a standardised data warehouse that holds genomic data and associated metadata. Any EGA submissions to date have been an ad hoc, manual process that involved submitting template CSV files through the EGA web-based interface.

Challenges encountered to date include:

- Difficulty understanding what the preferred and required metadata is for successful submission.

- Difficulty understanding exactly what files, checksums, and indices are required, resulting in back and forth communications with helpdesk and delayed submissions.
- Difficulty in assembling the metadata due to the lack of a standard metadata database to query.

ZERO/CCI

CCI has loaded 247 patient's data to the EGA for the Nat Med paper representing a cohort of rare kids' cancers. They have datasets for WGS (90x tumour, 30x normal), RNAseq, Methylation, CNV, SNV, SV under the study EGAS00001004572. They loaded FASTQ, VCF, text, and IDAT files and the process took 10 months to upload data due to capacity issues on the EGA side (limit uploads are capped at 10Tb per EGA box and you have to wait for them to archive the files before space is freed up to upload more files). Additionally, each submission was done in EGA's new portal via GUI, which was slow and not conducive to large uploads. The XMLs were relatively easier to automate. CCI has a bitbucket repo and confluence page for all steps. They used NCI to create the md5s and upload the data. The most confusing part was what to do if you had to re-use a subset of the same data for a different paper. For some other EGA Studies, they have had to upload the same files again.

Garvan Institute of Medical Research

The Garvan has had one cohort uploaded to the EGA (the Medical Genome Reference Bank, which consists of 4,000 samples). During the upload, there were capacity issues due to the shear size of the CRAM files. In addition, the amount of metadata that was required and the different XML schemas that were used were quite complicated and had several different tiers.

The Garvan does not use the EGA as a regular destination for data uploads due to these complications. Instead, different projects are scattered across different platforms (e.g., Garvan-owned hardware, NCI, and commercial cloud). These platforms each have their own challenges and downsides and are not a comprehensive replacement for EGA. Some of these challenges include but are not limited to: high storage costs, issues with data sharing, no metadata management, and no cataloguing of data or cohorts.

B. Archiving User Stories

User story	Role	Institute
As an Australian researcher, I want to access large international datasets quickly and easily	downloader	MISC
as a researcher, I want to easily share my data with collaborators after uploading to the EGA	submitter	Garvan
As a data submitter, I need to submit a dataset quickly and easily in order to get an accession to enable publication	submitter	QIMRB
As a data custodian, I want to enable easy access to disease specific sub-cohorts from a broader EGA dataset	custodian	ZERO/CCI
As a researcher, I want to deposit published data into a local EGA, with much faster TAT than to the Euro EGA	submitter	ZERO/CCI
As a data submitter would appreciate a more comprehensive list of metadata options to choose from and provide	submitter	QIMRB
As a researcher, I want to upload the data to the EGA with ease including my metadata.	submitter	Garvan
As a researcher I want to find other cohorts of interest and request access to them/download them with ease.	downloader	Garvan
An increase in quotas so that more datasets can be submitted for large projects	submitter	QIMRB
As a data engineer, I want an easy way to encrypt genomic data in the EGA format, in a way which encourages pushing data sample-by-sample, rather than batching all the encryption jobs first	submitter	ZERO/CCI
As a researcher, I want to compute and run pipelines on the data that I access on EGA.	downloader	Garvan
Clear transparency for process of when data is successfully archived or not	downloader	QIMRB
As a project PI, I want a solution that is easier/faster for collaborators to download approved data from EGA- usually downloads are cutoff and a pain	downloader	ZERO/CCI

User story	Role	Institute
As a data engineer, I want a convenient way to de-identify our genomic data, prior to uploading it to EGA. ie modify the header of BAM and VCF files	submitter	ZERO/CCI
As a data engineer, I want a convenient way to map from our data model to the EGA data model	submitter	ZERO/CCI
As a DAC, I would like a way to share a subset of a dataset to approved researchers	custodian	ZERO/CCI

C. FEGA Maturity Model

Full model is available at: https://inab.github.io/fega-mm/.

Below are the categorised indicators with the level required to become an operational node (Capella-Guitierrez 2022). Essential indicators must reach Level 4, important indicators Level 3, and useful indicators Level 2.

Essential Indicators

[1.1.1] Dedicated governance bodies and structure defined for the Federated EGA instance.

4. Governance body is fully operating with key personnel and is monitored based on work plan.

[1.3.1] Immediate resources (short-to-mid term funding)

4. Costed plan developed and implemented with some resources.

[2.1.1] Data Protection Impact Analysis (DPIA) performed

4. Considering the existing regulation and taking into account needs and proportionality aspects, propose measures to mitigate risks relevant to the Federated EGA Node.

[2.1.2] Risk and Vulnerability Analysis performed

4. Obtain approval of the risk management plan in accordance with the hosting institution of the Federated EGA node.

[2.2.1] Federated EGA Node Collaboration agreement established

4. Federated EGA Node collaboration agreement is signed and the Node becomes a recognized part of the Federated EGA.

[3.1.1] Secure installation for the Federated EGA node

4. Security policies and infrastructure are established and implemented under the appropriate jurisdictional level, e.g. nationally.

[3.1.4] Security breach/incident response plan defined

4. Production SOP in use, approved by relevant legal and/or security personnel.

Important Indicators

[1.3.2] Long-term sustainability

3. Available funding identified and stakeholders have been mandated.

[2.3.1] Data Processing Agreement (DPA) is available to users *

3. Data Processing Agreement template incorporates elements from the Federated EGA Ecosystem to ensure consistency.

[3.1.3] Risk register/assessment implemented

3. Risks identified, documented and assigned to appropriate personnel for review.

[3.2.1] Incoming Data flow in the federated EGA node is established

3. Ad-hoc incoming data flow into the Federated EGA node. This is a largely automated process.

[3.2.2] Outgoing Data flow in the federated EGA node is established

3. Ad-hoc distribution of data out of the Federated EGA node to approved users using labour intensive protocols.

[3.2.3] Mechanisms for sharing metadata and other operations-oriented information are established between the federated EGA node and Central EGA.

3. Communication interfaces are well defined. Information, operations-oriented & public metadata, e.g. study metadata, accessions, can be exchanged between the federated EGA node and Central EGA in a manual way.

[3.3.3] Metadata standards & harmonisation implemented

3. Metadata is ingested with limited formatting and minimum standards. Basic tools used for metadata collection (e.g. spreadsheets) and validation are deployed. Metadata management is partially automated.

[4.1.2] Interoperability with CEGA microservices implemented (e.g. permissions API, submission API).

3. Implementation and successful performance of minimally required communication with CEGA microservices.

[4.1.3] Microservices/APIs specific for FEGA node operations implemented.

3. Minimal set of APIs/microservices in production to support core FEGA node services. Additional APIs/microservices being developed/tested.

[4.2.1] Compliance testing

3. Implementation and successful performance of the core compliance tests as defined in the Federated EGA ecosystem.

[4.3.2] Storage Robustness

3. Node has a system to prevent data loss.

[4.4.2] Network Reliability / Security

3. Node has implementation of mitigation strategies for vulnerabilities (port security on switch, ARP certification, IP source guard, etc). An incident reporting system is drafted and partially implemented allowing to gain experience on those incidents.

[5.1.3] Helpdesk and SOPs established

3. Helpdesk working using FEGA SOPs.

[6.2.1] Training for the Federated EGA Node users

3. Training materials are available at the Federated EGA Ecosystem and mapped to the users.

Useful Indicators

[1.2.1] Roadmap/plan defined for the Federated EGA instance

2. Initial strategy/framework being drafted.

[1.4.1] Implementation, adoption and usage of KPIs in the Federated EGA node

2. Set of KPIs for overall implementation of the federated EGA node work plan as well as for the global and individual records usage are drafted taking as reference the ones from the federated EGA network.

[3.1.2] Data Access mechanisms are available in the Federated EGA node following Data Access Committee approval

2. Fully manual system. The process is triggered manually after approval is granted.

[3.3.1] Assessment of the Data Content Quality

2. Initial definition of the data content quality aspects to monitor based on existing agreements in the Federated EGA ecosystem.

[3.3.2] Community-agreed data standards and file types are implemented

2. Initial definition of the data standards and file types that will be supported by the federated EGA node in accordance to its mandate.

[4.1.1] Development Best practices

2. Identified best practices at the federated EGA.

[4.2.2] Stress Testing

2. The Federated EGA Node starts drafting stress tests based on existing knowledge at the Federated EGA ecosystem.

[4.3.1] Storage Capacity

2. Storage capacity needs are not planned but it is addressed ad hoc if the node has no more storage to provide.

[4.4.1] Available Network Capacity

2. Federated EGA Node network needs are covered by the hosting institution in an ad hoc basis.

[4.5.1] Available Computing Capacity

2. Internal computing capacity needs are not planned, but sufficient computing resources can be obtained as the node needs.

[5.1.1] Internal node operation SOPs defined and documentation available

2. Processes defined and documented in draft format.

[5.1.2] CEGA-FEGA interaction SOPs defined and documentation available

2. Cross processes identified and documented in draft format.

[5.2.1] Training and Capacity Building

2. The different needs of the Federated EGA node teams are assessed, gaps are identified and training options are under development.

[6.1.1] Awareness raising

2. A plan to engage users from the appropriate community starts in the dialogue on the importance of federated data sharing.

[6.3.1] Developed communication package

2. Initial (and informal) communications for the Federated EGA Node are happening.

D. Data Archiving Candidate Solutions Comparison Table

Requirements Identified as in scope	CEGA	FEGA	DBGaP	TDR
Capacity to load large projects/cohorts	soft limit 8TB	depends on implementation	UNKNOWN	No known limit
de-identify data	yes	yes	yes	user managed
Secure upload	encrypted + Aspera/FTP	encrypted + Aspera/FTP	yes	yes
Speed of data download/upload	no	yes (in Australian infra, more options)	no	yes (in Australian cloud instance)
Ease of data download/upload (user interface)	good enough - difficult learning curve	up to implementation. Potential to not download	UNKNOWN	No need to download, cloud download speeds
easy upload of metadata	difficult learning curve, can be overwhelming	depends on implementation	UNKNOWN	custom data model?
Findable data	Search/Browse catalogue	Search/Browse catalogue	Search/Browse catalogue	Search/Browse catalogue
accessible data	Controlled access	Controlled access	Controlled access	Controlled access
metadata model flexibility	metadata included in uploads, limited flexibility information	metadata included in uploads, limited flexibility information	metadata included in uploads, limited flexibility information	Metadata with uploads. flexible metadata schema
publication accession number/DOI	yes	yes	yes	UUIDs - are unique ids for datasets

usage cost	Free	depends on implementation	Free	Customer pays cloud usage
Requirements not in scope				
Metadata flexibility	no	no	UNKNOWN	yes
Compute near storage	no	depends on implementation	no	yes
share subset of a dataset	no	depends on implementation	no	yes
Other comments	This is a good and popular system overall other than the fact that it is difficult to upload to and download data from.	Many of the features would depend on how the FEGA is implemented.	A system much like the EGA, we were unable to find out some of the features for this platform	Seems like a robust and useful system. One drawback is the cost of using cloud based storage is left for the user to manage.
strength	Free indefinite storage for published research	Can be implemented to the required specifications and can remain on Australian hardware/cloud	Free indefinite storage for published research	"close to compute snapshot/ release concept"
weaknesses	Difficult to upload data to	N/A	Aimed at NIH funded studies.	Customer pays cloud usage