A Federated Data Platform to Drive Adoption of Advancing Technologies in Healthcare

Genivate23 :: Data registry :: CKAN

ABSTRACT

This innovative platform is poised to revolutionize clinical collaboration by facilitating seamless data sharing, enhancing access control through granular permissions, and enabling federated queries across diverse datasets, ultimately fostering a more integrated and responsive healthcare ecosystem.

Rust Turakulov

www.genivate23-data-registry.agrf.org.au

Genivate23 Project Report

01/11/2024

Project: A Federated Data Platform to Drive Adoption of Advancing Technologies in Healthcare

Team: Rust Turakulov, Application Developer (Bioinformatics); Lesley Gray, Data Strategy Manager; Matthew Tinning, Head of Laboratory Operations; Jacqueline Montgomery, Clinical Project Officer; John Fitzpatrick, Bioinformatics Support Specialist

Alliance Institution: AGRF

CONTENTS

BACKGROUND INFORMATION

Genomic technology plays a crucial role in modern healthcare, offering accurate personalized diagnoses, targeted treatments, and preventative strategies based on an individual's genetic profile. These innovations lead to better patient outcomes, reduced healthcare costs, and advancements in understanding complex diseases. However, the complexities of genomic diagnostics and medicine demand diverse platforms and assays, creating challenges for organizations striving to keep pace with technological advancements. High costs, limited access to cutting-edge technology, and the difficulty of obtaining suitable samples (particularly for rare diseases) are common hurdles faced by genomics organizations.

Data sharing presents a potential solution to these challenges, enabling collaborative efforts that accelerate innovation, reduce costs, and improve healthcare outcomes. A federated data platform can facilitate such collaboration by streamlining access to de-identified clinical data, integrating diverse datasets, and enabling comparative analysis. This approach fosters cooperation across institutions, promoting the adoption of advancing genomic technologies in healthcare.

USE CASE SCENARIOS

1. The gold standards data share for accreditation and validation purposes (Controls).

   Very often data with known properties or material source must be shared for validation purposes. The laboratory often required to do: cross platform sensitivity/specificity test; analytical pipeline changes or versioning controls, material processing kit performance monitoring. Those activities can benefit by utilizing and reusing same data (files) shared within and across organizations. Such sharing of control data can save time and money for replicating result and improve reproducibility of genomic assays.

   Examples:

   - Mock bacterial community sequencing data for 16s rDNA bacterial identification (Zymo Controls).
   - Somatic mutation DNA dilution experiment artificial mixture data. The FASTQ files for sensitivity test of the variant calling pipeline parameters and caller calibration.
   - Data for the sample with known chromosomal abnormalities or mutation for control purposes.
   - Virus or bacterial pathogen sequence or reference genome.
   - Mitochondrial genome assembly and/or raw data for mitochondrial disease studies.

2. Clinicians with interest in cohorting of rare cases (Multicentral cooperative data).

   One of common problem with studying rare disease cases is limited amount of cases to make statistical generalization. Putting demographics and metadata for the open access registry can help to establish potential collaboration with users who may have similar data for the chortling. Please note that actual genomic files do not have to be public. They can stay private until owner can approve sharing state with particular user or collaborator.

   Examples:

   - There are few rare sarcoma cases characterized with specific appearance on tissue slide which never been published and those cases shown characteristic gene fusion detected with RNAseq however there is only few cases per year with such disease coming through the pathology lab where those were observed. Potentially it can be new sarcoma type but to characterize it comprehensively it is requiring more samples and additional genomics tests to be performed. The Clinician can publish metadata and demographics for available dataset and mark them open for the collaboration efforts.
   - Two laboratories working together to collect and share rare disease cases with new tumour type. Both laboratories running complimentary technologies for different genomics tests. One laboratory has access to RNAseq assay and another laboratory providing proprietary in situ immunostaining microscopy for the same sample. Results for both tests can be shared with proposed platform and reduce data sharing coordination burden.

3. Long term project with incremental data and metadata accumulation (Data versioning).

   In project which are running for extended period of time the data management can be cumbersome. Proposed data registry platform can be used to preserve information about project samples over extended periods with controlled access to the metadata and raw files.

   Examples:

- Project data generation is scheduled over two years. The AGRF data policy specifies a three-year retention in the archive (this is an arbitrary number). By the end of the project, the first batch may only have one year left in the archive before being deleted. Moving the data to the data registry platform allows harmonization of metadata, management of the data lifecycle, and control over data availability across participating groups and vendors. In many cases, data should only be released to collaborators after the final batch is completed.
- In a similar situation, some samples may have been repeated at the beginning of the project, and then the same DNA was sequenced again at the end due to issues with DNA extraction. The platform allows for sample searching and data management, enabling the review and sorting of duplicates and repeats before the final release. This helps to manage multiple versions of incomplete data.

4. Data for share to collaborators prior to make it public (Controlled access).

   Collaborators may need to access the private data in organized way before data became available to public.

   Examples:

   - Project data needs to be shared with a postdoc researcher from the same or another institute. Access can be granted as an "Authenticated/Registered" user, allowing them to download data but not edit files or metadata.
   - Project data needs to be shared with a collaborator from another institute. Access can be granted as an "Editor" user, who can upload data (e.g., analysis files like a clinical report) and modify metadata, such as updating diagnosis fields or populating demographic entries. This user can also revoke access for other "Authenticated/Registered" users who may have moved out of the project or organization.
   - An organization has a large dataset, but only a few samples are of interest to a particular collaborator outside the organization. These specific samples can be shared with a user registered on the platform as an "Authenticated/Registered" user. This user will be able to see other samples and metadata within the project but will not have access to raw files for samples they are not authorized to view.

5. Prepublication file share for reviewer access (Controlled access).

   One common issue with making data public for journal publication is providing reviewer access to the data. Frequently, data uploaded to public repositories, such as the Sequence Read Archive (SRA) or the European Nucleotide Archive (ENA), has a fixed embargo period after which it automatically becomes public. This can make it challenging to reverse the release date in case of publication delays. The proposed platform offers greater flexibility for data sharing prior to publication.

   Example:

   - a project was prepared for publication, and data was shared with the journal reviewer through SRA. However, the review was negative, and the paper was resubmitted for another round. By the time the journal issued a second negative decision, the SRA embargo had expired, and the dataset became public. At that point, it became difficult to contact SRA, especially since the user who uploaded the files had changed their institutional affiliation.

6. Automatic data and metadata download/upload for routine processing clinical dataflow (File transfer and data pipelining).

   The platform offers API (Automation Programming Interface) to make programmatic connection to the data registry for uploading/downloading files and metadata modifications. Each user on platform has unique API key which allow access to user credentials and permitted dataset.

   Examples:

   - An ongoing project involves uploading and sharing fortnightly samples among users within the organization. The programmatic script can be launched manually by the operator on-premises or scheduled to run automatically, synchronizing files generated by the sequencing provider (AGRF) with the data registry. Files are uploaded to an S3 bucket associated with the data registry and can then be distributed by the dataset manager according to organizational rules. An API key for file uploads will be assigned to the dataset manager, thereby transferring the responsibility of file redistribution and end-user notifications from the sequencing provider to the data custodian.
   - Data stored on the data registry platform can be automatically downloaded using an API and a user key file generated on the platform. This key file inherits the user's access permissions to datasets, facilitating

automated downloads of new files whenever updates occur. Such scripts can be scheduled for daily checks for updates, which is especially beneficial for clinical data that requires a quick turnaround time.

- Additionally, bulk updates for metadata can be accomplished through the API, allowing for the generation and population of sample-specific metadata fields in bulk. For example, when mutation data for a project has been reviewed and annotated by a clinician or variant curator, the project can be reannotated based on confirmed pathogenic mutations.

7. Storage and archiving (long term storage)

The platform offers cost-effective data storage from both short- and long-term perspectives by allowing data to be registered and stored in independent AWS S3 buckets that utilize tiering across different classes based on access patterns. S3 provides several tiers, including S3 Standard for frequently accessed data, S3 Intelligent-Tiering for automatic cost optimization based on access frequency, and S3 Glacier or S3 Glacier Deep Archive for long-term, infrequently accessed data, which have lower costs but longer retrieval times. Users can efficiently manage costs over time by moving data between these tiers using lifecycle policies. Importantly, the S3 bucket can be fully managed by the data custodian or user, with the data registry platform requiring only the necessary credentials to access the user's S3 bucket. This setup enhances convenience for extracting subsets of data from the archive without the need to handle file permissions, create extra copies, or transfer large files unnecessarily.

Examples:

- A project involving exome sequencing data for an entire cohort of 400 samples was uploaded to a company-operated S3 bucket. The S3 storage is registered for Intelligent-Tiering with AWS to optimize long-term storage costs. While the S3 bucket has metadata assigned to all stored objects (fastq files), redistributing and subsetting this data for collaborators can be challenging. Additionally, accessing the data can be cumbersome for ordinary users who do not have an AWS account. The data registry platform offers a simple user interface that allows users to navigate across all samples and manually download subsets of 10 samples without requiring an AWS account or additional permissions for those files in the S3 bucket. Users can be added or removed from particular sets of files at any moment through the data registry platform.
- Another project is stored on a client's premises using an FTP server facility. In the same folder, there are data samples unrelated to sharing, which complicates access. Instead of creating copies of subsets of data for sharing and managing FTP usernames for access, this can be easily handled with the data sharing facility provided by the data editor. This eliminates the need to communicate with the system administrator for the FTP site.

USER BENEFITS

The proposed data registry platform offers several key benefits tailored to meet clinical requirements and data management needs. For users, the platform provides a secure and controlled environment for data sharing, ensuring that sensitive clinical data does not leave the premises. This is crucial for healthcare organizations bound by stringent privacy and compliance regulations. The platform enables seamless collaboration by granting access to de-identified datasets and metadata, allowing for clinical research, cohort studies, and validation projects without compromising data integrity or patient confidentiality. For data custodians, the platform's federated query capabilities allow users to search and compare metadata across multiple projects, streamlining workflows and avoiding duplication. This system not only improves access control and data management but also supports long-term storage solutions like AWS S3, ensuring scalability and efficient data lifecycle management for clinical and research purposes.

- **Secure data sharing:** Ensures clinical data remains on-premises, meeting privacy and regulatory requirements.
- **Controlled access:** Enables flexible user permissions, allowing de-identified data to be shared without compromising patient confidentiality.
- **Federated queries:** Facilitates the search and comparison of metadata across multiple projects, promoting collaboration and reducing duplication.
- **Efficient collaboration:** Simplifies sharing between institutions or teams by managing data access through roles and permissions.
- **Scalable storage:** Supports long-term data storage with AWS S3 integration, optimizing costs and ensuring easy data access.

- **Data lifecycle management:** Manages data archiving, retention, and controlled release, reducing the burden on users and custodians.
- **User-friendly interface:** Simplifies navigation and data retrieval without requiring advanced technical knowledge or AWS accounts.
- **Flexible API integration:** Enables automated data uploads/downloads for clinical workflows and programmatic data processing.

PLATFORM DESCRIPTION

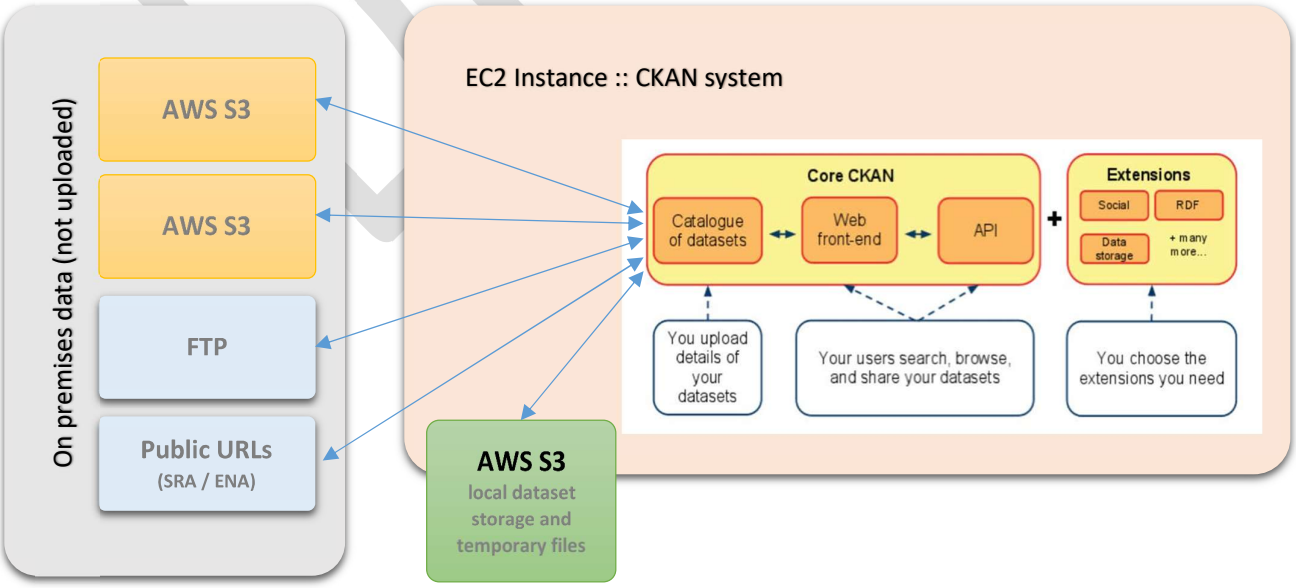
The data registry built around CKAN platform in core. CKAN is an open-source project that not only helps organizations publish and share their data but also enables more advanced functionalities, like managing data across multiple sources and controlling access at a detailed level. One key feature of CKAN is its ability to support federated queries. This means users can search and query datasets that are stored in multiple locations—whether in different departments, institutions, or even across geographic regions—without needing to move the data to a central location. CKAN acts as a bridge, allowing seamless access to distributed datasets, providing a unified view while keeping the data in its original storage.

In terms of data storage and access control, CKAN is flexible. Data can be stored on-premises, meaning organizations can keep sensitive information within their own servers while still making it available through the platform. This is particularly important for organizations with strict data privacy and security requirements. CKAN provides granular permission settings, allowing administrators to control exactly who can see, edit, or download certain datasets. Permissions can be assigned at different levels, from individual datasets to entire data repositories, and can be managed based on user roles or specific needs. This ensures that only authorized users can access sensitive or restricted data, while the general public or other stakeholders can still access open datasets.

In terms of functionality, CKAN not only helps you publish data, but also provides tools for data visualization and analysis. Users can create charts, graphs, or maps directly on the platform, making the data easier to understand. CKAN also supports metadata—information about the data—so users can know where the data came from, who published it, and when it was last updated. This makes CKAN an ideal platform for open data initiatives, government agencies, or any organization that needs to make its data available to a wide audience.

AWS AND SYSTEM ARCHITECTURE:

The CKAN system deployed on AWS EC2 instance and uses attached AWS S3 bucket to store local (uploaded) dataset and temporary files required for IGV browsing and normal CKAN work. This attached bucket configured in the way it can be only accessible from CKAN instance (green box). Usage of expandable S3 bucket is cost effective solution over provisioning fixed disk volume.



User files can be provided as direct upload to the system or alternatively as links to own S3 bucket or FTP sites those location access security credential will be stored in CKAN system configuration. Access will be managed with CKAN facilities for dataset user management.

The metadata fields, such as demographics or keywords will be indexed and discovered/searchable for all users however actual genomic files (BAM, VCF, can be only accessed by for user who can appropriate access level).

The CKAN API is a flexible interface that allows users and developers to interact programmatically with the CKAN platform, providing access to nearly all core functionality such as creating, updating, and searching datasets, managing metadata, and handling users, organizations, and resources. It is built on a RESTful design, enabling easy integration with external systems and automation of data workflows. Each user can generate an API key, which is required for authentication when performing write operations like creating or modifying datasets. This ensures controlled access and enables users to securely interact with CKAN, while also supporting the integration of external services or the automation of tasks.

Table with docker containers details used inside platform.

Container Name	Container Image name	Role
ckan	client-agrf-portal_ckan	This container is running the **CKAN** core application. This container provides user interface and web server application. The service HTTP requests exposed on port 5000. Provides API interface to user data. This container is linked to all other containerized services.
db	client-agrf-portal_db	This container is running **PostgreSQL**, the database backend used by CKAN to store metadata about datasets, users, and organizations. It's exposed on port 5432 for database queries.
client-agrf-portal_datapuser_1	ckan/ckan-base-datapusher:0.0.20	**DataPusher** service, a component of **CKAN** that processes uploaded BAM, VCF, CSV and Excel files, converting them into a form that can be previewed within CKAN's web interface including IGV. Additional feature is updates indexed metadata.
redis	redis:alpine	The **Redis** fast, in-memory key-value data store. Redis is used by **CKAN** for caching and improving the performance of the system for the search engine in session management environment. Every time a user logs in, their session information (such as authentication tokens and session state) is stored in Redis. Even if CKAN's web server is restarted, user sessions remain intact because Redis persists session data, ensuring users don't have to log in again.
solr	ckan/ckan-solr:2.9-solr9-vector	**Solr** a search platform integrated with **CKAN** to handle full-text search capabilities. It allows users to perform fast searches across datasets and metadata. CKAN uses Solr to index its dataset metadata. Every time a new dataset or resource is added to CKAN, its metadata is sent to Solr for indexing.

The CKAN system architecture for accessing on-premises datasets is designed with clinical data regulations in mind, ensuring that sensitive genomic data remains securely stored within controlled, on-premises environments. This setup prevents the generation of extra copies, safeguarding data integrity and minimizing the risk of data breaches. CKAN acts as a centralized platform where users can access metadata and request permission to view genomic datasets, but the actual data never leaves the organization's secure servers. Instead of transferring files, CKAN facilitates controlled access through granular permissions managed by the data owner, ensuring only authorized users can view or download the data.

For users with granted access, the platform provides integration with IGV (Integrated Genomics Viewer), enabling the visualization of BAM or VCF files directly from the system. This eliminates the need for users to manually download large files, as the data can be streamed securely through the platform while remaining within the institution's firewall. This

architecture meets stringent clinical data regulations by ensuring full control over who can access and view sensitive genomic information, while adhering to the requirement that no unnecessary data copies are generated.

The CKAN platform offers robust security features that have made it a trusted solution for many government agencies and institutions handling sensitive data. Its security architecture includes role-based access control (RBAC), allowing administrators to define specific permissions for users based on their roles, ensuring that only authorized individuals can view, edit, or manage datasets. CKAN also supports advanced authentication mechanisms like OAuth, LDAP, and integration with external identity providers, which ensures secure user verification and access management.

Additionally, CKAN is widely adopted by government organizations due to its compliance with stringent data security regulations. The platform offers encryption for data both at rest and in transit, providing protection from unauthorized access. Its audit logging capabilities ensure that every action on the platform is traceable, which is critical for meeting compliance with regulations such as HIPAA, GDPR, and national data security policies. This makes CKAN a reliable and secure choice for institutions needing to protect sensitive or regulated datasets while maintaining flexibility in how data is accessed and shared.

The system is structured to ensure that only users assigned to an organization and holding the "Data Editor" role can upload or link new datasets, preventing unsanctioned data storage. This controlled access ensures that only authorized personnel are responsible for adding or modifying data to the platform. Additionally, any data uploaded must be de-identified in compliance with HIPAA regulations, meaning no personally identifiable information (PII) should be stored in the system. However, the responsibility for ensuring that the data is properly de-identified rests solely with the data owner, who must ensure that the information adheres to all applicable privacy regulations before uploading.

CKAN PREBUILD PLUGINS AND OFF THE SHELF EXTENSIONS

CKAN project is well established has lot of open-source extensions that allow you to customize and enhance the functionality of a CKAN installation. CKAN has a modular architecture, meaning you can add, modify, or disable features without modifying the core codebase. Here is some examples which are not installed on MVP site.

- **Data visualizations**: Plugins like **recline_view** enable rich data previews with graphs, tables, or maps.
- **Harvesting**: The **ckanext-harvest** plugin allows CKAN to automatically pull data from external sources into CKAN's catalog. Allows CKAN to harvest datasets from external sources like other CKAN instances, government portals, or web services Automates importing and updating datasets.
- **Geospatial**: The **spatial** plugin enables geospatial metadata support and integrates CKAN with services like **GeoServer**.
- **R client for CKAN RESTful API: ckanr** is an R client for the CKAN API opens avenue to the Shiny server integration.
- **Authentication and Authorization**: Plugins for **single sign-on (SSO)**, OAuth, and more advanced user management systems. Or for another plugin like **ckanext-aaf** allows AAF (Australian Access Federation) authentication to log into a CKAN installation.
- **Many more on official page (>267)**: https://extensions.ckan.org/

CKAN plugins provide a structured means to extend and customize the core functionality of CKAN, allowing it to be tailored to specific requirements. Plugins enable enhancements to the user interface, integration with external services, and the development of custom data workflows, making CKAN adaptable to various contexts. With a wide range of pre-built plugins available, alongside the option to create bespoke extensions, CKAN can be transformed into a versatile, scalable, and inclusive platform for managing, integrating, and publishing data in a collaborative and efficient manner.

KEY FUNCTIONALY OF THE PLATFORM

USER ROLES MANAGEMENT AND ACCESS LEVEL.

User can do self-registration then contact data owner. There 4 tiers of users access to the platform and one system administrator level.

	User role	Access	Permission
1.	Anonymous / Unauthenticated nonregistered	Can view public datasets and resources.	No access to create or edit datasets or resources. Cannot access private datasets or perform administrative tasks.
2.	Authenticated / registered	Can view and search datasets and resources, including those that are private (if granted access).	Create Datasets: Can create and edit their own datasets and resources. Upload files. Edit Datasets: Can edit datasets they own. Collaborate: Can be added as a collaborator to other datasets. View Data: Can view and download datasets and resources they have permissions for. Create Organizations: Can create and manage organizations they belong to. Need to be approved for the organisation affiliation by Editor.
3.	Editor	Can view and manage datasets within their organization or those they have access to.	Edit Datasets: Can edit datasets they own and those shared with their organization. Manage Resources: Can add, edit, and delete resources within datasets. Approve Datasets: Can review and approve datasets if configured to do so. Collaborate: Can invite users to collaborate on datasets.
4.	Administrator	Full access to all datasets, resources, and organizations in CKAN.	Manage Users: Can create, edit, and delete user accounts. Manage Datasets: Can edit or delete any dataset and resource. Manage Organizations: Can create, edit, and delete organizations. Configure CKAN: Can change system-wide settings and manage CKAN plugins. Perform Administrative Tasks: Can manage data imports, site settings, and perform system-wide actions.
5	Sysadmin (System Administrator)	Superuser with unrestricted access to all areas of CKAN.	Full System Access: Can perform all actions of an Administrator. Manage System Settings: Can manage core system configurations and perform maintenance tasks. Access All Data: Can access and modify all datasets and resources. System Maintenance: Can handle tasks such as database backups, system upgrades, and configuration changes.

ORGANIZATIONS AND GROUPS CONCEPTS

Organizations primarily manage data internally, while groups can bring together datasets from multiple organizations. For example, in a research consortium, each partner (organization) may contribute datasets to a common research project (group). The group provides shared access to the datasets while each organization retains control over their own data. This setup supports secure, controlled data sharing and collaboration in CKAN, allowing flexibility in managing who sees and interacts with the data.

Organizations:

- Definition: An organization in CKAN represents a collection of users and datasets. It is typically used to group datasets and users that belong to the same entity, such as a department, institution, or company.
- Purpose for Data Sharing: Organizations help manage datasets and control access within a specific group of users. Datasets owned by an organization can be shared within the members of that organization or made public. This allows organizations to securely manage their data and control how it's shared or accessed externally.
- Access Control: Members of an organization can have different roles with varying permissions:
    - Admin: Can manage users, datasets, and resources within the organization.
    - Editor: Can edit and add datasets but may have limited access to administrative tasks.
    - Member: Can view the organization's datasets but has no editing or admin rights.
- Granular Permissions: Organizations allow for fine-grained control over who can access, view, or modify datasets. For example, some datasets can be kept private within the organization while others are shared publicly or with specific users.

Groups:

- Definition: A group in CKAN is a collection of datasets that share a common purpose or theme, regardless of which organization owns them. Groups can be created for specific projects, topics, or collaborations that span across multiple organizations.
- Purpose for Data Sharing: Groups are used to facilitate collaboration between different organizations or users working on a common project. Datasets from multiple organizations can be grouped together, making them easily accessible to collaborators.
- Access Control: Similar to organizations, groups can have specific access rules:
    - A group can be public, where anyone can see the datasets.
    - A group can be private, restricted to specific users, allowing for controlled data sharing in closed collaborations.
- Collaboration Tool: Groups are ideal for cross-organization collaborations where datasets from different sources are pooled together for a common purpose, such as a research project or an analysis task. Users from different organizations can access and contribute to the group's datasets without affecting the individual ownership of the datasets.

DATA PUBLISHING AND MANAGEMENT:

- **Data Storage**:
    - On-premises or cloud-based storage for datasets. Supports AWS S3 and SFTP. For complying with clinical security protocols genomic data can be kept on owner data premises without transferring to CKAN host facility.
    - Centralized or federated storage options with flexible infrastructure integration. Support multiple data sources for the same project or dataset.
- **Federated Queries**:

- Perform cross-database queries across multiple distributed data repositories. Search performed within indexed fields
- Unified access to datasets from various locations, with results presented in a consolidated format.
- Advanced search functionality for finding datasets by keywords, tags, or metadata.
- Support for faceted search and filters.

- **Data Publishing and Management**:
  - Easily upload and manage datasets through a web interface.
  - Metadata support for detailed dataset descriptions and categorization.
  - Automatic annotating genomic files in dataset with CSV provided metadata.
- **API Access**:
  - RESTful API for data access and automation.
  - Programmatic interaction with datasets for querying, retrieving, or updating data.
- **Collaboration Tools**:
  - Share datasets with specific users or groups.
  - Enable collaboration with version control and user-specific access.
- **Extensible Architecture**:
  - Modular plugin architecture to extend CKAN functionalities.
  - Integrations with third-party services and tools.
- **Data Visualization and Analysis**:
  - IGV plugin for checking BAM and VCF files/
  - Tools for creating visual representations of datasets.
  - Basic data analytics and export options for external analysis.
- **Data Provenance and Lineage**:
  - Track data source and changes over time.
  - Ensure transparency in how data is generated and modified.

LIMITATIONS

File Size Limit (10MB max for direct storage API upload)

Currently, the maximum file size for direct storage on CKAN's attached disk (S3 bucket) is 10GB. This limit has already been increased from the default CKAN configuration, which allows files up to 10MB, using a single-threaded upload protocol. Please note in this version we developed multithreaded plugin for the upload. For larger files, such as raw FASTQ files that typically exceed 10GB, it's recommended to upload them to a dedicated project S3 bucket and store them via API or by providing a remote URL. This bypasses the size restriction for direct file storage.

Search Limitations (Index-based search only)

At present, searches can only be performed through indexed fields—wildcard searches within stored files (e.g., text, CSV, PDF or other formats) are not supported. Search and filter functions work solely on metadata entered by users into indexed fields or tags for each file or dataset. Implementing full-text search across all stored files would require significant development, potentially using a Lambda function to generate temporary summaries once a search is completed.

User Access List Not Centrally Available

The complete list of users with access to a dataset is not currently consolidated into a single view. As a workaround, you can retrieve the list of users with organizational access and their roles, along with a separate page for collaborators who can access individual datasets (including external collaborators not included on organization). This information can also be indirectly managed or extracted using the "Audit log" or by developing a custom extension that logs dataset access events in more detail. Another way to manage dataset access is "Groups" tab where user will be able to add and see each other organized by the datasets of interest.

SUMMARY

Our concept delivers an EOI, pilot data register and minimal federated data access mechanism based on live, *in situ* AGRF clinical datasets. Through this platform, the Alliance would leverage reference data, gold standard datasets, and share clinical insights from emerging sequencing technologies and bioinformatics workflows. By first offering an environment to share this data, the Alliance can collaborate on generating the rules and standards around datasets, aiming to establish conventions that build consensus and allow seamless integration across systems. If successful, this allows for tools and workflows to be built collaboratively, and to expand the platform for sharing and running pipelines among the Alliance.

While some elements of this concept have been explored before, our approach distinguishes itself through a unique combination of data management process and cutting-edge technology – a winning combination for genomic information management.

REFERENCES:

MVP site:
https://genivate23-data-registry.agrf.org.au/

GitHUB documentation:
https://github.com/datopian/client-agrf-portal

https://github.com/AGRF/clinical_data_registry

CKAN project homepage and documentation:
https://docs.ckan.org/en/2.11/

CKAN Extension list (official)
https://extensions.ckan.org/

Consulting contractor – Datopian
https://www.datopian.com/