Aridhia FAIR Data Services Version 1.1.7

Service description

Author: ***Gideon Giacomelli***
Date: ***April 2021***
Version: ***1.1.7***

Table of Contents

Executive Summary

Aridhia FAIR Data Services gives researchers and innovators the ability to discover and understand data through dataset search, classification and efficient metadata browsing capabilities described via dataset catalogues, dictionaries and associated attached assets. Researchers can request access to datasets where data owners can review and either approve or deny where approval unlocks an action such as data delivery to an Aridhia Workspace.

FAIR Data Services is designed for research and to help reduce the barriers to entry of discovering and browsing data – tasks that researchers can spend up to 80% of their time performing to facilitate their research. FAIR Data Services also provides a **secure** and **compliant** environment that is designed with security and privacy in mind, prevents unauthorised access or use of data and adheres to **information governance** standards.

Researchers and innovators have the option to integrate with Workspaces to run analysis and curate new data in a secure and audited environment, which can then be **re-published to** enhance data findability.

Scope of document

This document describes the features and functionality of the Aridhia FAIR Data Services product.

Aridhia DRE Overview

Aridhia provides a Digital Research Environment (DRE) to enable the collation, analysis and sharing of healthcare research data in a collaborative manner. This platform allows a research project or organisation to securely upload and store pseudonymised data in a cloud hosted environment which can be accessed by collaborators across the globe. The data can easily be searched and collected for analysis in the Workspace.

The DRE comprises two main components, FAIR Data Services and Workspaces, facilitating the following workflow in Figure 1. This document focuses on the FAIR Data Services. Information about Workspaces can be found in the Workspaces Service Description.

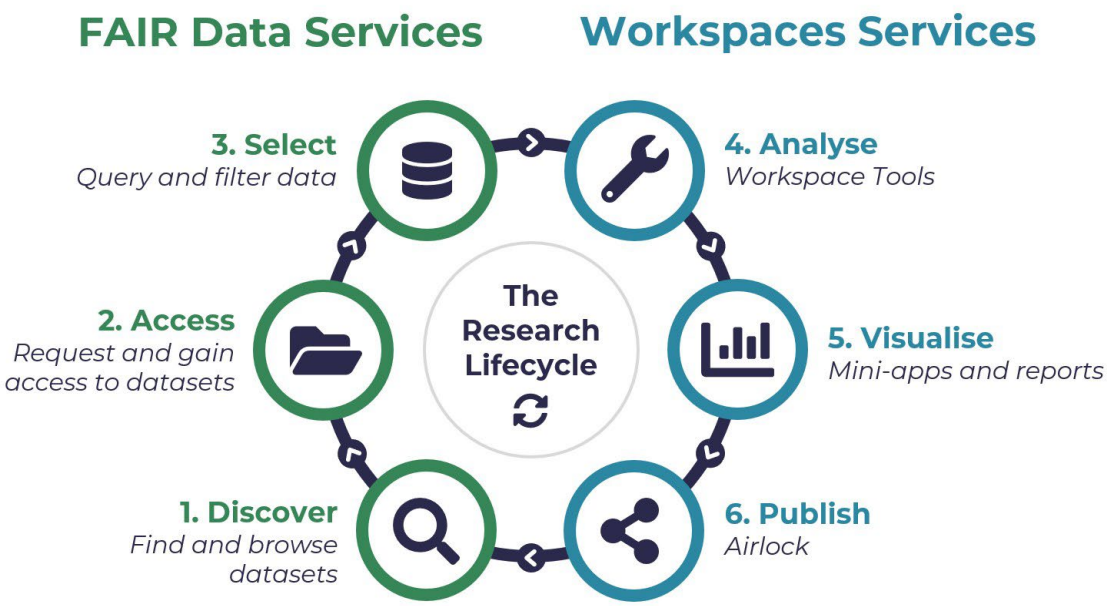

Figure 1: The Research Lifecycle

FAIR Data Services Features

Key Features

DATA DISCOVERY

- Search for datasets relevant to your research project using text-based simple or complex search queries.

METADATA BROWSING

- Understand existing datasets by viewing metadata including catalogue and field-level descriptions
- Download machine-readable dataset metadata.

METADATA MANAGEMENT

- Upload your dataset metadata and associated attachments (e.g. PDFs, json, etc) to be discovered by others.

REQUEST ACCESS TO DATA

- Users can request access to datasets to unlock a pre-defined action (e.g. delivery to a Workspace).
- Data owner review and decision of user access requests.
- Data owner assignment of a request workflow per-dataset.
- Custom access request forms

ROLE-BASED ACCESS CONTROL

- Self-service signup with role-based user permissions. This includes managed and self-service custom roles.

BUILT ON STANDARDS

- Uses the Data Catalog Vocabulary (DCAT) for dataset instance-level descriptions.

INTEGRATION WITH ARIDHIA WORKSPACES

- Single Sign On (SSO) between FAIR and Workspace services
- Consistent Aridhia DRE user interface.

PRIVACY BY DESIGN

- Secure data access and management via MFA, RBAC, encryption and secure key management
- ISO 27001 accredited.

CLOUD-NATIVE SERVICE

- Developed and hosted on the cloud.
- Integrates with and improves on cloud technologies.

This document also outlines the technical aspects of the service such as deployment options, service-level agreements (SLAs) and security as well as the process of Onboarding the customer to FAIR Data Services.

Information Architecture

This section describes and defines the information architecture in relation to data that can be hosted by the service.

A dataset entry consists of the following entities:

- A single **catalogue** entry describing the dataset at the instance-level that conforms to the DCAT standard.
- One or more **dictionary** entries describing the dataset at the field-level. This includes:
    - dictionary description
    - field name
    - field label
    - field type
    - description of the field
- **Controlled vocabularies** describing a field's set of constrained values
- One or more **attachments** that will enhance the metadata of the dataset such as PDFs, JSON, etc
- A Dataset Access Request (DAR) workflow allowing the dataset to become requestable (optional).

A dataset in FAIR Data Services is defined as a single entry in the FAIR service aligned to a single dataset catalogue, i.e. a dataset is equivalent to one dataset catalogue including its attachments and associated data dictionaries.

Service Overview

Data Discovery

Users can discover datasets in FAIR Data Services via the search function. The FAIR search is available via the FAIR user interface and API and supports both simple and complex Lucene-based queries.

The search service currently indexes and searches over the following entities:

- Dataset catalogues conforming to the Data Catalog Vocabulary (DCAT) standard.
- Dataset dictionaries including table names, field names, field labels, field types, field descriptions and per-field controlled vocabularies.

A successful search returns a result list of matching or similar datasets outlining:

- The total number of datasets returned.
- Dataset name (i.e. catalogue heading).
- Dataset description with highlighted matching terms.
- Dataset dictionary fields.

Searches can be saved and re-run at a later date via the in-built saved searches functionality or the URL of the search can be copied and either bookmarked or shared externally with other users of the service. Note that if the underlying dataset has changed, the execution of existing saved searches may return different results.

Metadata Management

Users are able to create and manage their datasets via the FAIR user interface and API.

The dataset creation process allows users with the appropriate permissions to:

- Add a dataset **catalogue entry** conforming to the DCAT standard.
- Add one or many **data dictionaries** describing the dataset at the field-level.
- Add one or many **controlled vocabularies** per dataset that can then be assigned to dictionary fields.
- Add **attachments** that will enhance the metadata of the dataset or are relevant to the dataset. For example, JSON files, PDFs, etc. The number and size of resources are subject to limits specified in the service fair usage policy.
- Self-**classify** the dataset via the use of keywords.
- Assign a DAR **workflow**.
- Modify the **visibility** of the dataset. Currently there are two levels of visibility:
    - Visible to me (i.e. private): the dataset is private to the user that created it.
    - Visible to all (i.e. internal): the dataset is visible to all users within FAIR Data Services.

Note that only the data owner (i.e. the creator) of a dataset can only edit or delete datasets they have created.

The list of all datasets within the service alongside their description is viewable in the 'All Datasets' tab.

FAIR Data Services will support a fair usage limit of up to 200 datasets. If this soft limit is reached, Aridhia and the customer will discuss ongoing requirements.

Metadata Browsing

Users are able to browse selected datasets and view associated metadata to help further understand the purpose and relevance of the datasets to requirements. The metadata available to view is as follows:

- A dataset catalogue entry describing the dataset at a high level including:
    - Dataset name (mandatory)
    - Dataset description
    - Author and Author Email
    - Publisher
    - License
    - Version
    - Rights
    - DOI
    - Keywords
- One or more data dictionaries (e.g. for multiple database tables) describing the fields of the dataset including the field names, types and descriptions as well as their controlled vocabularies.
- Attached resource files such as JSON, PDF, etc that can be downloaded to view their contents.

Catalogue and dictionary entries can be downloaded from the user interface or API in DCAT-compatible JSON and JSON formats respectively.

Requesting Access to Data

Data Access Requests (DAR) allow users to request access to datasets in FAIR. A Data Access Request is composed of entities.

1. The required information to be supplied by the requestor. FAIR Data Services allows custom DAR forms to be set per DAR workflow (see below) or Aridhia can provide a pre-defined template with the following information:
    a. Request name
    b. Project name

c. Project description
d. Project end date
e. Purpose
f. Pubic interest rationale
g. Required tables
h. Destination Workspace (if data delivery is required)
i. Destination Workspace Hub (if data delivery is required)
j. Applicant name
k. Applicant position
l. Applicant telephone number
m. Applicant address
n. Confirmation of dataset terms and conditions

2. A DAR workflow defining the stages in the DAR process. Note, FAIR Data Services by default only supports one default DAR workflow that allows for request, review and decision. More complex workflows can be created upon request to Aridhia.

The user shall assign a DAR workflow and one or more data dictionaries to the dataset to make is requestable.

Data Access Requests requested by users are sent to the data owner(s) of the dataset who can review. The supported decisions are supported:

- **Pending:** awaiting review by the data owner(s).
- **Approved:** approved by the data owner(s) to access.
- **Denied:** denied by the dataset owner(s).

Email notifications are sent to the data owners(s) when a request has been made and to data requestors when a decision has been made regarding the request.

Request decisions must be supplemented with a reason for the decision that is reflected back to the requestor.

Request creation, approval or denial and deletion actions are audited and are retrievable via the API with appropriate permissions.

Signup and Role-Based Access Control

In order to use the FAIR service, all users are required to signup. The signup process is self-service and users can do so by navigating to the URL provided by Aridhia, for example: **https://fair.<customer>.aridhia.io**. The signup process is described on the Aridhia Knowledge Base, however this signup process requires:

- Email address and password for login and user verification purposes.
- Mobile number for Multi-factor Authentication (MFA) login.
- First and last name used to display throughout the portal (e.g. logged-in user, audit events)
- Job title.

A user newly signed up to the service will be assigned a role (default: Observer) however the user will remain unapproved until an Administrator approves the user. These roles can be assigned to users by the Customer in a self-service manner at any time or by Aridhia upon request. Similarly, the service can be configured to auto-approve users after signup upon request.

A set of managed roles exist that can be assigned to users:

Role	View Datasets	Request Access to Datasets	Add/Delete Datasets	User and Service Management
Observer	☑			
Standard	☑	☑		
Data Steward	☑	☑	☑	
Administrator				☑

- **Observer**: a role that only allows the user to view existing datasets made public by Data Stewards
- **Standard**: a role that allows the user to view and request access to existing datasets made public by Data Stewards but can also save searches.
- **Data Steward**: can create, edit and delete datasets as well as assume the permissions of the Standard role.
- **Administrator**: can perform role management capabilities and other administration tasks via the API.

Where managed roles differ to requirements, **Administrators** can create custom roles from the FAIR Data Services permission set and assign these roles to users in the platform. Role creation and deletion actions are audited and are retrievable via the API with appropriate permissions.

Service Administration

FAIR Data Services can be administered by the Customer (with appropriate permissions). The following administrative functions exist:

- **Roles**: the management of roles including the assignment of roles to users and the creation of custom roles. These roles can be composed of permissions that are available in the FAIR Permission Set.
- **Configuration Vocabulary**: configuration of links outwith the service. Three vocabularies and their links can be modified both via the UI and API:
    - **Terms and conditions**: modify where to direct users to view custom terms and conditions
    - **Service logo**: modify where to direct users to when the service logo (top left) is selected
    - **Link dropdown**: add a custom dropdown of links that users can select and be directed to. For example, a list of different Workspaces.
    - **Platform name**: modify the default platform name listed on the home page (default: Aridhia DRE)
    - **Data Access Request Emails**: the contents for the DAR email notifications can be modified for the request, approval and denial emails. Further email addresses can also be added as CC or BCC.
- **Themes**: set a service theme from a list of those provided. Management via the UI and API.
- **Users**: view a list of registered users, their assigned role and whether they are approved or not.
- **Audit**: view the service audit displaying the who, what and when of audited events in the service. These events are currently:
    - **Datasets**: uploading and downloading of dataset attachments
    - **Data Access Requests**
        - Creating and deleting of Data Access Requests
        - Approving or denying a Data Access Request
        - Updating a Data Access Request Workflow
    - **Role Management**
        - Creating or deleting a Role
        - Updating a role's permissions
    - **Users**: approving or unapproving a user from the service.
    - **Configuration Vocabulary:** updating a vocabulary.

User Interface

A user logged in to FAIR Data Services with the appropriate permissions will be navigated to the home page. The home page provides an overall summary of the service including:

- A top-level status summary including the total number of datasets available to search and view.
- Featured datasets.
- Service-wide recently uploaded datasets.

The home page provides links to search functionality, the listing of existing datasets and the ability to find out more about FAIR Data Principles and FAIR Data Services. Note, users assigned to different roles will see differing versions of the user interface dependent on their assigned privileges.

The user interface layout can be modified by following the instructions described here.

The user interface supports theming where a service administrator can set a service-wide theme or allow users to select their own theme from a pre-defined enabled list. The user interface's themes can be configured by following the instruction described here.

API

FAIR Data Services is developed using an API-driven approach, therefore all features and operations available via the user interface are available to perform via the API (if not more). The API specification can be found at the */api/docs* path after your FAIR instance URL, e.g. **https://fair.<customer>.aridhia.io/api/docs**.

Technical Information

Deployment Options

FAIR Data Services is developed upon and hosted on Microsoft Azure. Users can select to deploy the service in an Azure Region of their choice. Typically, this decision is made based on the Azure Region closest to the user's locale or that satisfies their data governance requirements. Note that onboarding into some Azure Regions may not be possible due to a Region's technical constraints. While this is extremely rare, this may occur for example, if the Region does not support a feature our service requires.

FAIR Data Services is hosted by Aridhia on a customer dedicated Azure Subscription. Customers requiring to host their own FAIR Data Services deployment on their own tenancy, can be considered.

The resource footprint of the service deployment will be discussed with the customer during the onboarding process. Any resource footprint changes to be made during the operation of the service, either requested by Aridhia or the customer, shall be agreed between both parties before actioning.

AZURE SERVICES USED

FAIR Data Services currently uses following Azure services within a FAIR-specific resource group:

- Virtual Machine (and associated resources such as Disks, Networking, etc)
- Storage Account
- Azure Search Service
- Azure Kubernetes
- Container Instances

- Key Vault
- Log Analytics Workspace

Service Level Agreement

Aridhia offers availability of 99.5% excluding planned downtime. Specific service performance SLAs are contractual and differ between clients.

- Support SLAs are described in the Aridhia Service Desk SLA document. SLAs may be customised to customer requirements at Aridhia's discretion upon request.
- Terms and conditions of the service can be found in the End User License Agreement (EUA).

Security

Aridhia has implemented the DRE (and therefore FAIR Data Services) to ISO 27001 standards certified by external auditors. This covers Aridhia's processes and technology (see certificate here). The DRE is deployed on Microsoft Azure which is also accredited to ISO 27001 as well as many other environmental, systems management and cloud standards.

ENCRYPTION AND KEY MANAGEMENT

Aridhia follows best practice guidelines for data encryption and secure key management for all cryptographic operations:

- All user and API communication is encrypted using HTTPS.
- Internal resource passwords and keys are rotated on a regular basis in-line with best practices.
- Currently dataset metadata held within the service is encrypted at rest using FIPS 140-2 compliant 256-bit AES encryption.

MFA AND ROLE-BASED ACCESS

Multi-Factor Authentication (MFA) is enforced for all user accounts of the service. The principle of least privilege is employed via role-based user access, particularly when newly signed up users are given no permissions until assigned by a FAIR Administrator.

USER INTERFACE

- The current list of supported browsers are described in the Aridhia DRE Browser Support Knowledge base article.
- The Mozilla Observatory Score for the FAIR Data Services user interface is an **A+.**

Onboarding Process

In order to commence onboarding, Aridhia will conduct an onboarding session with the customer that will cover:

- An overview of the FAIR Data Services onboarding process.
- Information gathering from the customer to assist in the onboarding of FAIR Data Services.
- Discuss and agree the deployment resource footprint based on cost, performance and scalability requirements.
- Providing information on the service SLAs and how the customer can access the Aridhia Service Desk.
- Providing training on how to use the service.
- Obtaining agreement from the customer on the onboarding process and estimated onboarding timelines.

Shared Responsibilities

This section outlines the responsibilities for Aridhia and the customer in order to successfully deploy and operate the service.

The customer shall:

- Submit any change requests, feature requests, incidents or general support questions to the Aridhia Service Desk.
- Identify a customer service owner(s) for the purpose of being a primary point of contact for the service.
- Identify a customer service administrator for the purposes of user management as described above.

Aridhia shall:

- Deploy the service and necessary infrastructure on an Aridhia customer-dedicated or customer's own Azure Subscription.
- Provide training to the customer on how to utilise the service.
- Provide support to the customer via the Aridhia Service Desk adhering to defined support SLAs.
- Continue to provide updates to the service where necessary (e.g. features, security) and notify the customer of such updates and any impacts to the running of the service (e.g. downtime).
- Act only as a processor of data and **not** a data owner.