NATIONAL AG-DATA HUB PROJECT TECHNICAL REPORT: (PHASE ONE)

A report from
ACATECH TECHNOLOGY
Arada Sub City,
Addis Ababa, Ethiopia
Email: info@acatechte.com
Phone: / (+251)91263-22-22/91189-60-97
https://acatechte.com/

ACATECH

October, 2022

AICCRA
Alliance Bioversity International CIAT
CIMMYT International Maize and Wheat Improvement Center
ICARDA Science for resilient livelihoods in dry areas
MINISTRY OF AGRICULTURE

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The Accelerating Impact of CGIAR Climate Research for Africa (AICCRA) project is supported by a grant from the International Development Association (IDA) of the World Bank. IDA helps the world's poorest countries by providing grants and low to zero-interest loans for projects and programs that boost economic growth, reduce poverty, and improve poor people's lives. IDA is one of the largest sources of assistance for the world's 76 poorest countries, 39 of which are in Africa. Annual IDA commitments have averaged about $21 billion over circa 2017-2020, with approximately 61 percent going to Africa.

EXECUTIVE SUMMARY

This document is a technical progress report of a work aimed to develop a national ag data hub in Ethiopia. The document is not a full technical implementation report of the national ag data hub. It is the technical report of the first phase as per the agreement entered between CIAT and ACATECH TECHNOLOGY PLC to implement the system in a phase by phase approach. This report has two main parts. The first part provides background information about the project which includes a brief introduction about the first phase deliverables. This shows the scope of the project to be completed in the first phase (i.e. in the last three months). The second part describes the project activities and progress in contrast to the plan set at the beginning. This part clearly showed the progress achieved so far. The following are some of the deliverables that are briefly described in the second part of the document.

1. The requirement specification document refinement is completed and a final requirement specification document is presented.
2. Detailed solution architecture based on microservice framework is designed, technology choices are decided.
3. Two development/test servers, databases and microservice are configured.
4. Installation and configuration of the version control system, source code repository, issue tracking, application containerization, integration and build automation tool, inter service communications among services are completed.
5. The first phase of metadata and data standards adaptation and implementation of these standards within the core data management service and the multimodal database is completed.
6. The first phase of the data hub information architecture and implementation of this architecture in the ag datahub system is completed.
7. The first phase of data harvesting pipeline implementation is completed. Permanent data harvesting links are implemented for more than 7 systems, more than 100 datasets each having many files.
8. The implementation of the first phase user interfaces/experiences is completed (i.e. home page, core data management, Admin panel dashboard page, User authentication and authorization, data harvesting interfaces, data Curation interface, Data visualization and reporting interface, query interface).

PART ONE: BACKGROUND OF THE PROJECT

1. INTRODUCTION

The national digital transformation strategy of Ethiopia considered agriculture among the sectors that need a special implementation strategy. As a result of this, initiatives to transform the Ethiopian digital agriculture system such as "National Digital Agricultural Extension and Advisory Services Roadmap 2030" are started. The main recommendation in such strategies is strengthening the national agricultural data system. The agriculture sector is mostly data intensive and digital transformation of the sector presupposes improving its data system. The national ag-data hub project is initiated by taking the above urgencies to strengthen data systems in the sector. The main goal of the project was developing a national ag data hub in Ethiopia that will help to bring independent data silos from disparate agricultural sources. The requirement and architecture of the ag-data hub was co-designed in collaboration with key stakeholders of the system from various government agencies working to improve the agricultural sector (i.e., The Ministry of Agriculture (MoA), Ethiopian Institute of Agricultural Research (EIAR), Agricultural transformation institute (ATI), National Metrology Agency (NMA), Geospatial Information Institute (GII), Ministry of Innovation (MINT) and so on. We applied the philosophy of co-designing and co-creation of system architecture for a sustainable ag-data hub in Ethiopia.

This document presents the technical accomplishments and progress done in the three months reporting period of the National Ag data hub project that was signed between Alliance Biodiversity CIAT and ACATECH Technology P.LC. In order to give a brief overview of the main project deliverables for interested readers, the objectives of the project as it is planned at the beginning of the reporting period is presented first followed by detailed discussion of the deliverables achieved in this phase.

2. OBJECTIVE OF THE PROJECT

The main objectives of phase one of this project is building foundational bases to build a national ag data hub. Achieving the above objective expects achieving the following specific objectives:

- Refining the software requirement specification document by conducting additional requirements study and modeling this using system analysis tools
- Designing a detailed Solution Architecture and deciding technology choices that will be used in the implementation of the agri-datahub.
- Configuration of the software development and testing servers, code repository and team collaboration platforms and project management platforms.
- Adapting metadata and data standards for core data management and data integration services and preparing them for further feedback to be improved in the remaining phases.
- Designing the Information Architecture of the national agri data hub system
- Implementation of the existing metadata and data standard as well as Information Architecture of the system.
- Designing initial user interfaces and assessing user experience in using such systems
- Implementing data harvesting pipelines for data sources having APIs and from structured sources that are producing data in the Ministry of Agriculture.

PART TWO: REPORT PROGRESS

2.1. SOFTWARE REQUIREMENT SPECIFICATION (SRS) DOCUMENT

In this section, the result of the requirement modeling is briefly documented. Though the software development approach is agile and users may not expect a well-documented list of requirements, a high-level requirement specification document that shows the rationale, scope, the methodology to be followed, the nature of the product is developed following the IEEE requirement specification guideline.

2.1.1. INTRODUCTION

Digital Ethiopia 2025 prioritizes the digital transformation of the agriculture sector and gives directions to initiate a digital transformation plan contextualized to the sector. One of the problems that are given attention in the 2030 roadmap is the establishment of a national ag datahub platform that will bring together the disconnected agricultural datasets into a single central source. There is a serious challenge related to the availability of quality and standard data in an organized manner. As a result, creating a seamless integration of datasets from different systems is complex due to the lack of common standards and formats. For instance, various organizations and stakeholders in the country have been collecting voluminous data related to agriculture, the environment, and infrastructure. ATI collects data on soils. EIAR has been collecting data related to agricultural experiments since the 1960s. The Ministry of Agriculture (MoA) has been collecting data on natural resources and agriculture for more than 50 years. Crop yield and related data have been collected by the Central Statistical Agency (CSA) for many decades. On the other hand, the National Meteorological Authority is the custodian of time-series and predicted climate/weather data. The Geospatial Information Institute (GII) holds tremendous aerial photographs and satellite data. There are also various other sectors and authorities, initiatives, projects, and individuals who have collected various datasets related to different domains.

Different sectors in the country have been investing millions of dollars for data collection purposes . However, the absence of standards has made access and integration of the data very difficult. The problem is so widespread that data sharing under even the same directorate has been a challenge. For instance, the National Soil Information System (NSIS) at the Soil Resource Information and

Mapping Directorate and the National Rural Land Administration Information System (NRLAIS) at the Rural Land Administration and Use Directorate, which are found at the MoA, are not yet integrated and cannot share dataset efficiently.

Efforts have been made to harmonize datasets to facilitate storage and data sharing. However, some of those efforts have already failed while some others are still struggling to succeed due to different reasons including lack of champion institutions and limited buy-in and the capacity of the government to support initiatives, limited cooperation between sectors and among Directorates within sectors, low technological readiness, governance barriers, and data incompatibility. Hence, developing a national datahub to bring these disconnected sources of information is a timely initiative. However, the development of such a system requires careful preparation due to challenges such as the complexity of the system, the early maturity stage of datahub technologies, lack of legal frameworks and standards, and shortage of human capacity

2.1.2. PROJECT SCOPE

The scope of the project can be grouped in three parts; development of the national Ag-datahub and regional Mini-datahubs, development of the Ag-data standards and best practices and capacity building. The national Ag-datahub is required to be a platform that integrates and consolidates the dispersed data in various local repositories, legacy systems, information systems, and cloud-based systems, and provides access to well-organized agricultural data. It is expected to assist in decision-making by policymakers at the federal and regional levels, experts and researchers in the agricultural domain, ag-extension workers, farmers, actors in the ag-value chain, and other local and international data users. The Ag-data governance and policy will be a document that facilitates the implementation of the Ag-datahub and guides its data-sharing features. The capacity development aspect of the scope is aimed at providing training and capacity-building programs to relevant stakeholders on the Ag datahub use. It is also expected to contribute to creating awareness and fill the skill gap in the area of data selection, preparation and publishing, data sharing policy, and governance.

The following are the functional requirements that fall under the scope of the Ag Datahub:

- Data extraction from source systems. Depending on the source systems the extraction will involve direct database access or API-based access.

- Triggering scheduled or on-demand data extraction jobs from the source systems.
- Facilitating full or incremental data extraction mechanism from a source system.
- Transformation of heterogeneous (i.e., format, structure, standard) data ingested from multiple data source systems to fit the target data system.
- Harvesting data from min-datahubs into the national ag-datahub. Mini-datahubs will be developed at selected institutions, such as MoA and the regional agriculture bureau.
- Integrating data of multiple source systems. This requires cleaning, normalizing, and harmonizing multiple or related data maintained about the same object.
- Providing flexible ag-data metadata management mechanism.
- Sharing of data from local repositories into the national Ag.datahub with appropriate access privileges.
- Contributing data to the national Ag-datahub.
- Downloading data from the Ag-datahub with appropriate access control. Download options could be direct or via API.
- Supporting data browse, search, view, and exploration on the Ag-datahub with proper access control.
- Querying the Ag-datahub with SQL or SPARQL command.
- Visualizing data using appropriate visualization tools.
- Verifying data quality for accuracy and completeness.

2.1.3. SOFTWARE DEVELOPMENT METHODOLOGY

The general project development workflow of the national ag- data hub development is shown in figure 1.

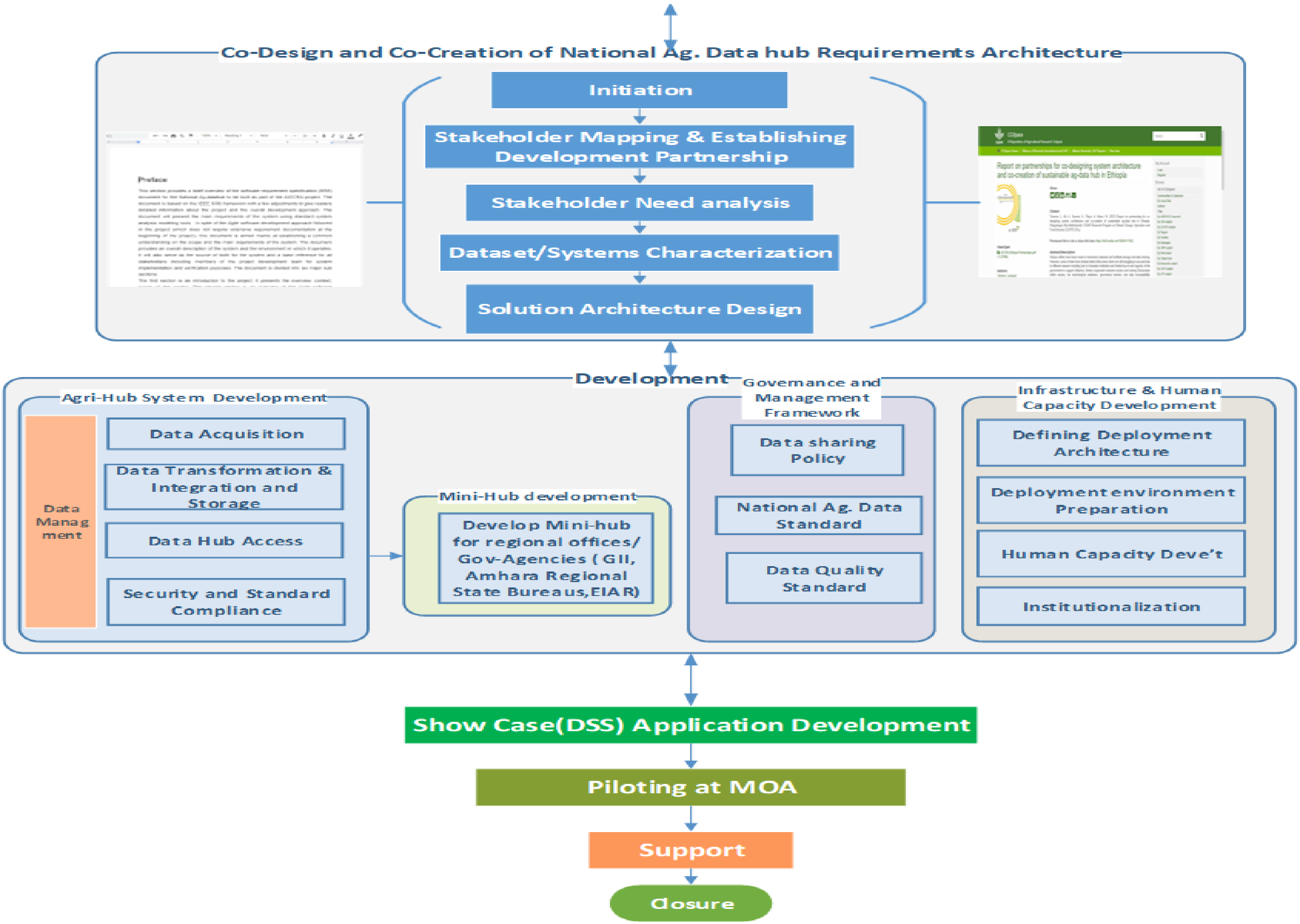

Figure 1:Ag-Datahub Development Workflow

The development of a complex system like that of the national Ag-datahub requires a careful selection of the appropriate software development methodology. Making the right decision regarding what software development methodology to use is not easy due to the several contextual factors affecting the decision. Whether a project is delivered within schedule, within budget, and meeting all the project requirements is a function of the choice of development methodology. Two groups of software development methodologies are discussed in comparison for the sake of clarification. The first is the waterfall method which is basically based on a sequential series of steps with very specific activities performed and deliverables produced at each stage. The main stages in the waterfall method, presented in sequential order, are requirements, design, implementation/coding, testing, and maintenance. The traditional waterfall method requires that one stage is complete and finalized before the project progresses to the next stage. However, the assumption that requirements can be fully understood until the project is into development and test stage is unrealistic and its practical adoption is less. The waterfall method might work well in a structured organization where there are formal approvals and milestones. It is also considered appropriate in a context in which the requirement is well-defined at the beginning and does not call for a change of requirement later in the development process. The nature of the Ag-datahub project, in this regard, does not fit well with the assumptions of the waterfall approach.

The philosophy of the Agile methodology, as it is articulated in the Agile manifesto (Beck 2001), advocates the importance of individuals, including focusing on strong working relationships among the team members and the need for the team and individual motivation. Agile methods focus on the minimization of risk by developing small segments of the project using iterative techniques. Each iteration should last up to four weeks at the most. Agile project planning within these iterations is designed to be team-focused and collaborative. The project planning involves the entire team and all team members must be actively involved in the project scheduling and priority analysis. Agile methods accept change and can support changes to the product even quite late in the development cycle. Therefore, they can handle projects where requirements are not well defined or prone to change. Moreover, Agile methods can minimize some risks faced by larger projects as they force the project team to break the project down into smaller, more manageable segments.

The Ag-datahub is a complex and multi-stakeholder system that has cross-organizational and sectoral nature. It is, therefore, impractical to have each and every requirement of the system collected and documented fully at the beginning of the project as the Waterfall approach requires. It is rather more effective and feasible to have the high-level overall scope and requirement of the system established, followed by an incremental development process in which each individual feature is addressed one at a time following its own iterative process. Thus, the project team selected the Agile methodology as it is found to be more aligned with the context and nature of the Ag-datahub project than the traditional Waterfall approach.

In order to develop each module or feature of the system using the Agile methodology, the development team:

1. Collect and model requirements together with project sponsor and main users of the system such as ALLIANCE BIOVERSITY CIAT, MOA, EIAR, ATI, etc.
2. Design and develop a functional software module based on the defined/baselined requirements
3. Test the module
4. Integrate and deliver the functional iteration/module into the production system
5. Receive feedback from owners, accept the module, and factor this into the requirements of the next iteration

In the spirit of the Agile philosophy, the Ag-datahub development process will be participatory in that it will engage key individuals from the MoA, EiAR, GII, CIAT, and other relevant stakeholders.

Review of MOA Systems: More than fourteen (14) information systems that are used to automate the tasks of different Directorates and projects are hosted at the Ministry's data center. The systems are web-based and store and manage different types of data (structured, unstructured, geospatial, etc.), and can be accessed by a variety of users in different parts of the country. The systems also capture real-time/near real-time data - for instance, Animal Disease Notification Information System (ADNIS). Different types of databases namely MongoDB, Mysql, PostgreSQL, and MS SQL Server are used to manage the data. In addition, PostGIS extension is employed to support spatial data management. API (application programming interface) is available with an opportunity to integrate data of individual systems. A wide variety of open-source and commercial technology stacks were used for developing the systems hosted at MoA. Java, JavaScript, Python, and C# are the main programming languages used, while, asp.net core, laravel, life-ray, asp.net MVC, ReactJS, and AngularJS are the major frameworks used for building the systems. The systems are deployed on-premise at MoA data center using virtualization technology - an abstraction of layers over computer hardware that allows the hardware elements of a single computer processors, memory, storage and more to be divided into multiple virtual computers, commonly called virtual machines (VMs) with Linux and Windows operating systems.

Review of ATI Systems: ATI has its own data center for hosting systems to manage various types of data and automate its day-to-day activities. Data acquired from different project activities in the form of excel files, MS Access tables, spatial data (vector and raster), and the unstructured PDF data available in the organization are directly imported into ATA's data hub. The major DBMS software used include MongoDB, PostgreSQL with PostGIS extension, and Mysql. PostgreSQL DBMS with PostGIS extension is used for the data hub. The systems and the data hub are deployed using the Linux operating system. Python and PG loaders are used to integrate systems, and PowerBI is used for visualization/dashboard purposes.

Review of EIAR Systems: The EIAR is the custodian of a large dataset related to soils/agronomy, crops, breeding, feed/forage, and NRM that have been collected since the 1960s. The recent collaboration with the Alliance of Biodiversity and CIAT under the CoW also attracted a big dataset associated with crop response to fertilizer application. In addition, EIAR and its partners have developed various systems and decision support tools such as the Ethiopian Digital Agro-advisory Platform (EDACaP), wheat rust surveillance system, and others.

Review of GII Systems: Currently GII does not have its own data center for managing its various systems. However, it manages the national spatial data infrastructure installed in a server at its main office. Since it collects, manages, standardizes, monitors, and shares different geospatial and associated data (metadata, attributes) of the country, it has a large volume of geospatial data (both vector and raster) in digital and analog forms.

Review of the FarmStack: FarmStack is an open-source decentralized data sharing protocol, which fosters coordination across the agricultural ecosystem while protecting privacy, security, and control over how data is used. Rather than requiring organizations to share their data with a third party, data providers and data consumers can use FarmStack's self-service peer-to-peer connectors by installing a certified software package on their existing infrastructure. FarmStack does not, however, support some of the key features of a functional data hub as intended by the AICCRA project.

2.1.4. PRODUCT SPECIFICATION

The national Ag-datahub is a system that integrates and makes data of disparate systems accessible from a central location. The data it integrates come from different government agencies that work in the agriculture domain at the federal and regional levels. The context diagram shown in Figure 2 shows the exchange of data between the Ag-datahub and its external entities.

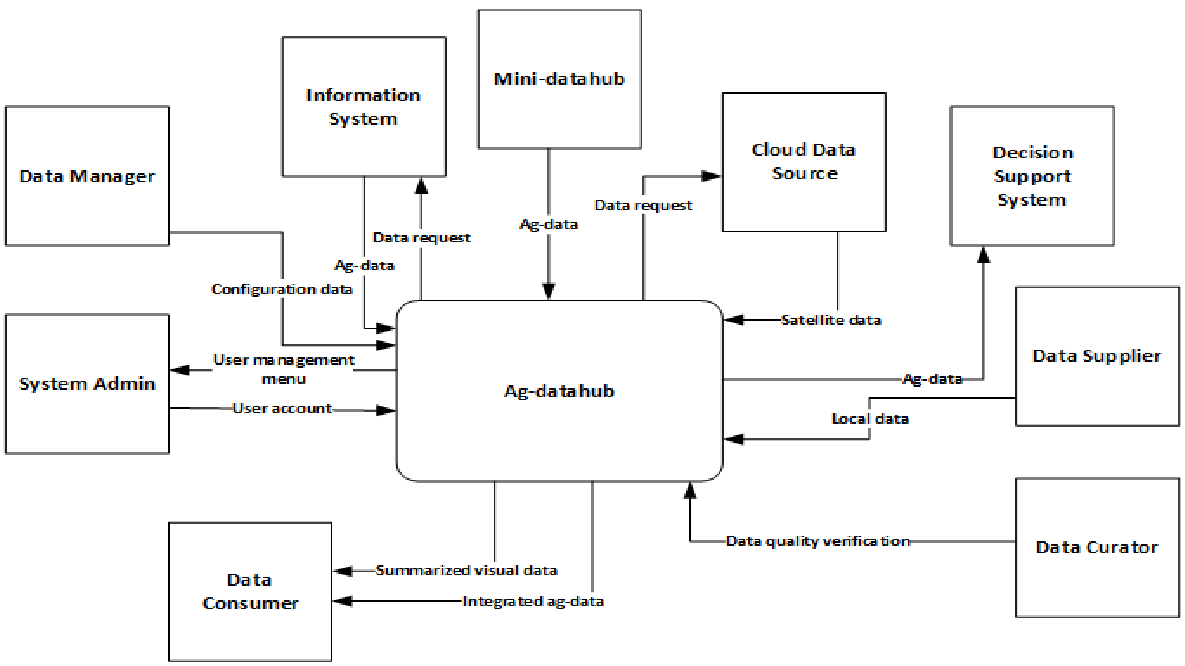

Figure 2:Ag-Datahub Context Diagram

Data is ingested from three main source systems; information systems, cloud data sources, and local legacy databases. In the case of the first two sources, the extraction shall be semi-automated. The data manager sets the data extraction and harvesting configurations with parameters including schedules and source system addresses. Data extracted from the source systems go through various transformation and cleaning processes such as normalization, merging, union, standardization, anonymization, annotation and harmonization. This process requires a temporary working area called staging storage. Once the transformation process is completed, the clean data will be loaded to the target storage of the Ag-datahub. Moreover, data located at local offline sources can be contributed by data suppliers. These data should be verified for their quality and correctness by the data curator. Data loaded to Ag-datahub will be used by data consumers based on their access privileges. The various decision support systems (DSS) that need the Ag-datahub as their data source will access the data through API and use them for their own data analytics purposes and support for decision making. The external entities and their descriptions are presented in Table 1.

Table 1: Ag-Datahub External Entities

External entities	Descriptions
System admin	Administrates user accounts and other system administration activities.
Data consumer	Users who access data from the Ag-datahub
Data Supplier	Contributors of data from offline sources into the Ag-datahub.
Information System	Various information systems that are accessed by the Ag-datahub for data extraction purposes.
Mini-datahub	Datahubs that will be deployed at selected key stakeholders for data management at the regional and federal levels.
Cloud data source	Raw data extracted from the cloud to be integrated to the Ag-datahub
Data manager	Is an expert responsible for setting data extraction and harvesting configurations on the Ag-datahub.
Data curator	This is an expert responsible for verifying the quality and correctness of data uploaded by data suppliers.
DSS	External systems that access data from the Ag-datahub for generating insights to support decision making.

2.1.5. OPERATING ENVIRONMENT

MoA is the owner and will host the national Ag-datahub. The Ag-datahub will, therefore, be deployed on MoA's infrastructure. The following table presents the operating environment and its specifications for the Ag-datahub.

Table 2: Operating Environment

Criteria	Operating environment
Server-side operating system	Debian-based operating system (such as Ubuntu 18.04 or above), and other operating systems (RedHat, Fedora, CentOS, OS X)
Web browser	• Google Chrome: Latest + previous version • Mozilla Firefox: Latest + previous version • Microsoft Edge
Web server	Tomcat, Apache HTTP Server (HTTPD)
Network /access for users	Internet connectivity
Server	Number & Speed of Processor: 16 core CPUS each with >2Ghz System RAM: >= 70 GB (for DB size >=4TB) Database size: >= 4 TB Storage size: 1.5 tera
Data storage/ database	PostgreSQL, Multimodal (document, key-value, graph) DBs
Datahub Search platform	Solr
User geographical location	Regional, federal, or international
Server's hosting institution	MoA
Location of databases	MoA's infrastructure

2.1.6. ASSUMPTIONS AND DEPENDENCIES

Assumptions are aspects of the project that are expected to be there for smooth functioning of the project. Dependencies on the other hand refer to the conditions that should be fulfilled as a condition for the system.

The following are the assumptions and dependencies of the system:

- There will be a single- contact person for decision making and approval of issues pertinent to the development of the Ag-datahub.
- Owners of the data at source system will collaborate with the development team by providing the required information and resources for data access and extraction.
- Data will be available for extraction and harvesting on the source systems found at the MoA and other pertinent ministries.
- Data at the collaborating ministries will be accessible to the Ag-datahub through one or more of the following options:
    - API
    - ETL/ELT Pipelines
    - Direct database access
    - Direct upload (Conventional File Manager System) through data-suppliers
- The required hardware infrastructure will be made available at MoA for the piloting and complete deployment of the system. The required software such as web server, database server will also be readily available.

2.1.7. SYSTEM FEATURES

The features of the Ag-datahub, as seen from the actors/external entities of the system, are presented in the form of a use case diagram as shown in figure 3.

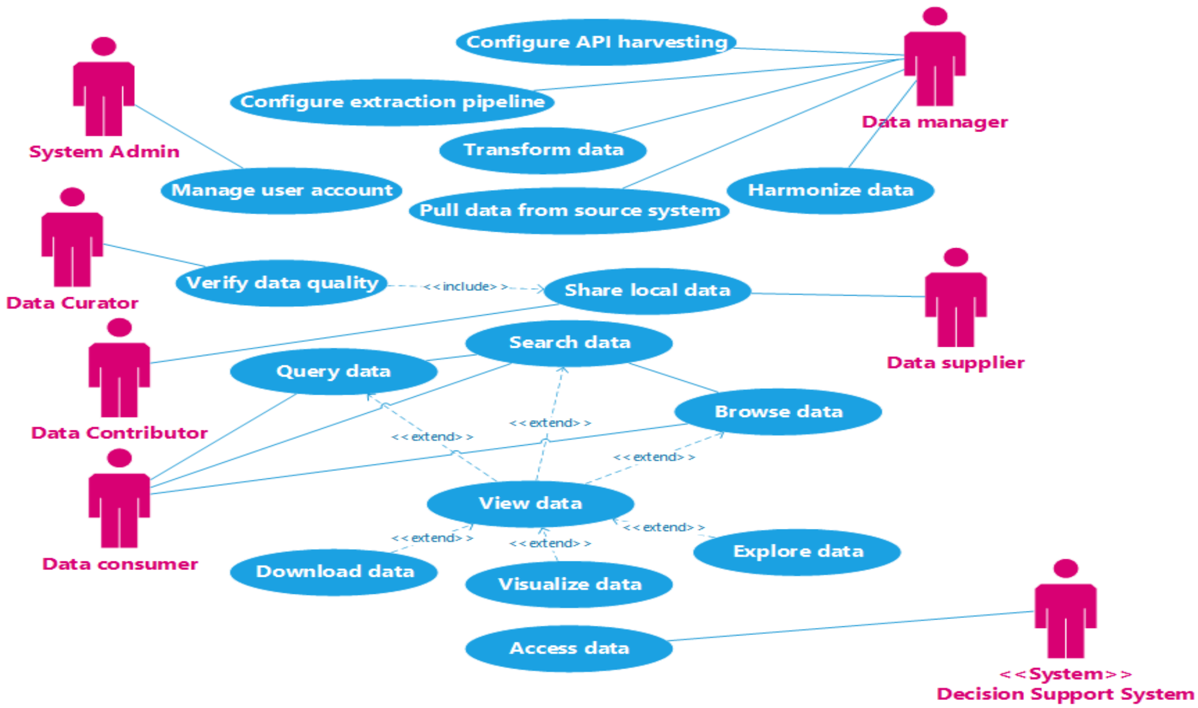

Figure 3:Use case diagram

The descriptions of all of the use cases shown in the above use case diagram are presented below.

Table 3: Use Case Descriptions

ID	UC-1
Name	Configure API harvesting
Include	None
Extend	None
Actor	Data Manager
Description	The data manager provides information about the source system, the type of data, harvesting mode (incremental or full) and frequency of harvesting.
Precondition	• The data manager is authenticated • The data manager is authorized to do the configuration • The source system is accessible online through API
Postcondition	• The configuration is set
Normal Flow	1. The data manager sets the configuration information such as the source system address, harvesting schedules, and the like in the API harvesting configuration form 2. The system validates if configuration information is complete and correct 3. The system notifies the data manager that the configuration information is set successfully

	4. The use case ends.
Alternative course of action	2a. The identifies an error in the configuration information 1. The system displays an error message that the configuration information is not correct 2. The system asks the data manager to enter the configuration information again 3. The use case continues at step 1 of the normal flow

ID	UC-2
Name	Configure extraction pipeline
Include	None
Extend	None
Actor	Data Manager
Description	The data manager provides information about the source system, the type of data, extraction mode (incremental or full), and frequency of extraction.
Precondition	• The data manager is authenticated • The data manager is authorized to do the configuration • The source system is accessible online through API
Postcondition	• The configuration is set • The data is harvested from the source system based on the set schedule
Normal Flow	1. The data manager sets the configuration information such as the source system address, extraction schedule, and the like in the extraction pipeline configuration form 2. The system validates if the configuration information is complete and correct 3. The system notifies the data manager that the configuration information is set successfully 4. The use case ends.
Alternative course of action	2a. The system identifies an error in the configuration information 1. The system displays an error message that the configuration information is not correct 2. The system asks the data manager to enter the configuration information again 3. The use case continues at step 1 of the normal flow

ID	UC-3
Name	Pull data from source
Include	None
Extend	None
Actor	Data Manager
Description	The data manager pulls data from source systems based on predefined configuration

	information.
Precondition	• The data manager is authenticated • The configuration to pull data is already set • The source system is accessible online through API
Postcondition	Data is pulled from source systems and loaded in the Ag-datahub
Normal Flow	1. The data manager opens the data pulling form 2. The data manager selects the desired configuration from the list of configurations 3. The data manager selects the type of data pulling (Extraction or harvesting) 4. The data manager selects the mode of pulling (Incremental or Full) 5. The data manager clicks on the pull button 6. The system starts the process of pulling the data 7. The system displays completion information 8. The use ends
Alternative course of action:	2a. No configuration is set 1. The system displays that no configuration information is set. 2. The system informs the data manager to set a configuration 3. The use case ends 7a. Failure to complete the pulling process 1. The system fails to complete the data pulling process 2. The system notifies the data manager to try again later. 3. The use case ends

ID	UC-4
Name	Transform data
Include	None
Extend	None
Actor	Data Manager
Description	The data manager transforms the loaded data through the process of data cleaning, normalization, and the like.
Precondition	• The data manager is authenticated • Data is loaded from a source system
Postcondition	• Data is transformed into the required form
Normal Flow	1. The data manager selects the data to be transformed 2. The system displays the data with options to select and change individual elements of the data 3. The data manager selects the data item and transformation type, and clicks okay. 4. The system confirms that the transformation is completed 5. The use case ends.
Alternative course of	4a. The system fails to do the transformation

action	1. The system displays an error message that the transformation failed, and informs the user to try again later 2. The use case ends

ID	UC-5
Name	Harmonize data
Include	None
Extend	None
Actor	Data Manager
Description	The data manager harmonizes data based on a predefined metadata of the Ag-datahub.
Precondition	The data manager is authenticated
Postcondition	Data is harmonized based on predefined metadata
Normal Flow	1. The data manager selects the data to be harmonized 2. The system displays the data with options to map with the predefined metadata 3. The data manager selects the harmonization option and clicks okay. 4. The system confirms that the harmonization is completed 5. The use case ends.
Alternative course of action	4a. The system fails to do the harmonization 1. The system displays an error message that the harmonization failed, and notifies the user to try again later 2. The use case ends

ID	UC-6
Name	Manage user account
Include	None
Extend	None
Actor	System Admin
Description	The system admin creates, updates, deletes, or sets access privileges to different types of user accounts.
Precondition	The data manager is authenticated
Postcondition	User account change is made based on the selected user account operation
Normal Flow	1. The system admin selects the account management link. 2. The system displays the user account management interface

	3. The system admin selects the option to create a new user account 4. The system displays the interface to enter user account details 5. The system admin fills the user account details and clicks okay to create the account 6. The system creates the account and displays a confirmation message 7. The use case ends
Alternative course of actions	3a. System admin selects the option to update a user account 1. The system displays the content of the selected user account 2. The system admin makes the required changes including change of details and access privilege types 3. The system admin clicks update 4. The system completes the update and displays a confirmation message 5. The use case ends 3b. System admin selects the option to delete a user account 1. The system displays a confirmation message to delete the selected account 2. The system admin confirms the deletion 3. The system deletes the user account 4. The use case ends 5a. System admin enters wrong data 1. The system displays an error message and notifies the user to enter the correct data 2. The use case resumes from where it stopped

ID	UC-7
Name	Share local data
Include	None
Extend	None
Actor	Data Supplier
Description	The data supplier provides complete metadata information about a resource and uploads the associated file to the system.
Precondition	The user is authenticated
Postcondition	The local data is shared in the Ag-datahub
Normal Flow	1. The user clicks the share resources link 2. The system displays the interface to share local data resources 3. The user enters details of the resource 4. The user uploads the associated resource 5. The system validates the completeness and correctness of the data 6. The system makes the resource ready for verification by the data curator 7. The use case ends.
Alternative course of actions	5a. The system identifies that there is an error in the data entered 1. The system displays an error message and notifies the user to enter the correct data 2. The use case resumes at step 3

ID	UC-8
Name	Verify data quality
Include	Share local data
Extend	None
Actor	Data Curator
Description	The data curator verifies the accuracy and completeness of the data shared by a data supplier for approval for publishing.
Precondition	There is data shared ready for verification
Postcondition	Data is verified and is approved for publishing
Normal Flow	1. The data curator selects a link to list resources shared by a data supplier 2. The system displays list of resources for verification and approval purposes 3. The data curator selects a resource and clicks okay 4. The system displays the complete details of the resource 5. The data curator verifies the resource for correctness and completeness. 6. The data curator approves the resource for publishing 7. The system publishes and makes the resource ready for access 8. The use case ends.
Alternative course of actions	6a. Data curator rejects the resource 1. The data curator verifies that the resource is not up to the standard for publishing 2. The data curator rejects the resource 3. The system removes the resource and sends a message to the respective data supplier about the decision 4. The use case ends 6b. Data curator suspends publishing 1. The data curator decides that the data requires corrections before publishing 2. The data curator clicks on the comment link 3. The system displays the comment form 4. The data curator enters the comment to the data supplier and clicks send 5. The system sends the comment to the data supplier to do the necessary corrections 6. The use case ends

ID	UC-9
Name	Search data
Include	None
Extend	View Data
Actor	Data consumer

Description	The system searches and displays data based on the search keyword the user provides.
Precondition	None
Postcondition	Relevant search result is displayed to the user
Normal Flow	1. The user enters the search keyword 2. The system runs Solr to search for the relevant data from its database 3. The system displays the result to the user in the order of relevance 4. The use case ends
Alternative course of actions	2a. No data is entered into the search input box 1. The system notifies the user that no data is entered and asks the user to enter data 2. The use case continues at step 2 of the Normal Flow. 3a. No related data is found 1. The system displays that there is no related data to the search keyword, and informs the user to enter another keyword 2. The use case continues at step 2 of the normal flow

ID	UC-10
Name	Query data
Include	None
Extend	View data
Actor	Data Consumer
Description	The system searches for data based on SPARQL or API query and displays the result according to the query provided by the user.
Precondition	None
Postcondition	The result of the query is displayed to user
Normal Flow	1. The data consumer selects SPARQL to query data in the Ag-datahub 2. The system displays the input box for accepting the SPARQL query 3. The data consumer enters the SPARQL query 4. The system analyzes the query and searches the relevant SPARQL end-points. 5. The system displays the result to the user 6. The use case ends
Alternative course of actions	1a. The data consumer selects API query 3. The system displays the form to enter the API query 4. The data consumer enters the API query and clicks okay 5. The system displays the result based on the query 6. The use case ends 2a. No data is entered into the query form box 1. The system notifies the user that no data is entered and asks the user to enter data 2. The use case continues from where stopped

ID	UC-11
Name	Browse data
Include	None
Extend	View data
Actor	Data Consumer
Description	The data consumer browses data in the system based on their alphabetical order, category, tags, and the like.
Precondition	None
Postcondition	The data consumer gets information about the available list of data
Normal Flow	1. The user selects a specific browsing option (Alphabetic order, Category, Tags, etc) 2. The system displays the data based on the selected browsing option 3. The user browses the data 4. The use case ends
Alternative course of action	2a. The system fails to display data based on the selected browsing option. (For example, no data may be available under a specific category) 1. The system notifies the data consumer that there is no data under the selected category, and informs the user to select other browsing options 2. The use case ends

ID	UC-12
Name	View data
Include	None
Extend	Explore data, download data, visualize data
Actor	Data consumer
Description	The system displays complete detail of a selected data
Precondition	None
Postcondition	Relevant data is displayed to the data consumer
Normal Flow	1. The data consumer clicks on the link to a dataset 2. The system displayed complete detail of the data 3. The use case ends
Alternative course of action	2a. The system fails to displays the detail of the data 1. The system displays an error message and notifies the user to contact the system admin 3. The use case ends

ID	UC-13
Name	Visualize data
Include	None
Extend	None
Actor	Data consumer
Description	The system displays an infographics or map view of a selected data
Precondition	The data has to be available
Postcondition	Relevant data is visualized
Normal Flow	1. The data consumer selects the link to visualize the data in the form of a map 2. The system displays a geospatial representation of the data. 3. The use case ends
Alternative course of action	1a. The user selects infographic visualization 1. The system displays the data in the form of bars, graphs, and other available forms of presentations 2. The use case ends 2a. The system fails to visualize the data 1. The system notifies the user that there is a problem with the visualization and informs the user to contact the system admin 2. The use case ends.

ID	UC-14
Name	Download data
Include	None
Extend	None
Actor	Data consumer
Description	The data consumer downloads data from the Ag-datahub to the user's local device
Precondition	The data must be available in the system in a downloadable form
Postcondition	Data is downloaded
Normal Flow	1. The data consumer selects the download link of the required data 2. The system checks the availability of the resource for download 3. The system downloads the data to the user's computer 4. The use case ends
Alternative course of action	2a. The data is unavailable for download 1. The system notifies the user that the required data is not available for

	download, and informs the user to contact the system admin 2. The use case ends

ID	UC-15
Name	Explore data
Include	None
Extend	None
Actor	Data consumer
Description	The user explores data based on exploration parameters
Precondition	None
Postcondition	The selected data is explored
Normal Flow	1. The data consumer selects exploration parameters of the data 2. The system lets the user explore the data based on the selected exploration parameters 3. The use case ends
Alternative course of action	2a. The system fails to allow the exploration 1. The system displays an error message, and informs the data consumer to contact the system admin 2. The use case ends

ID	UC-16
Name	Access data
Include	None
Extend	None
Actor	Decision Support System
Description	The system allows Decision Support Systems to access data of the Ag-datahub through API
Precondition	None
Postcondition	Data is accessed by DSS through API
Normal Flow	1. The relevant Decision Support System sends a request to access data through API 2. The system analyzes the request and responds with the required data 3. The use case ends
Alternative course of action	2a. The system finds an error in the API request 1. The system sends an error message to the requesting Decision Support System

	2. The use case ends

2.1.8. DATA REQUIREMENTS

Data requirements describe the data used by the national Ag-datahub system. The system is expected to host a range of agricultural data including crops and livestock. Based on the findings obtained from the different workshops, meetings, focus group discussions, and systems and data analysis activities conducted in collaboration with experts from participant organizations (MoA, EIAR, GII, ATA), a general category of Agri dataset is prepared as shown in Table 2 presented below.

Experts from multiple organizations were brought together in four different workshops[1]: two in Adama and two in Addis Ababa organized to produce a catalog of the systems and datasets of the represented organizations. Through a detailed analysis of the presentations and reports from the main stakeholders, nearly 20 categories of data have been identified for incorporation in the national Ag-datahub. The general list of agricultural dataset categories presented in Table 3 will be further reorganized based on more information that will be obtained later before implementation in the national Ag-datahub. Furthermore, a detailed metadata standard will be worked out in partnership with the relevant ministries and government agencies to maintain interoperability and successful integration of the system.

Table 4: Agricultural Datasets and Their Descriptions

No	Name	Description
1.	Spatial data	Thematic spatial data with varying coverage in the country (E.g., road network, infrastructure, etc.)
2.	Satellite imagery data	Real-time satellite imagery data coming from about 7 different satellites.
3.	Employment (farmer, animal herder, semi pastoralist)	The number of employment opportunities created for people in the agriculture sector.
4.	Exchanges (Markets)	This is information about the distributed/traded agricultural products
5.	Product prices	This is the local price of Agri-products in the country
6.	Agribusiness investments	This is data about the agricultural business investment activities

[1] *Refer to the methodology section of this document for the details on the workshop requirement gathering method*

		in the country.
7.	Hydrology data (Surface and groundwater data)	Spatial and non-spatial data about the availability and location of water resources
8.	Agro-climate	This is historical meteorological data on climate
9.	Animal production	This is the data about livestock information in the country.
10.	Animal export	This is data about the export of both live and slaughtered animals
11.	Animal diseases monitoring	This is about animal health monitoring data, vaccination, etc. in the country.
12.	Crop pest and disease monitoring	This is about monitoring data on major crops (maize, wheat, barley, maize, sorghum) in Ethiopia
13.	Soil	This is information about the soil characteristics of crop lands in the country.
14.	Agri inputs (fertilizers, seed, animal food)	This is data about the import and distribution, and application of Agri inputs in Ethiopia.
15.	Crop production	This data is about the production details of crops in Ethiopia.
16.	Land use and land productivity	This is information about the productivity of crop lands in the country.
17.	Advisory (extension) services	This is about the types and number of advisory services offered to farmers by government and private agencies.
18.	Forestry	This data is about the forest coverage in Ethiopia.
19.	Knowledge production (publications, policy, strategy, guides, etc.)	This data is related to knowledge production in the Agri-sector in Ethiopia.

2.1.9. CONCEPTUAL DATA MODEL

A conceptual data model is a representation of concepts and ideas that form the services of a system. In other words, conceptual data models represent concepts (entities) and their relationships, portraying the business domain at a high level, using non-technical terms and concepts familiar to the business users. This high-level representation enables effective communication among stakeholders and the technical developers on how the data aspect of the system is going to be realized.

Having identified the types of potential data to be stored in the national Ag-datahub, an overarching conceptual data model that will support the types of Agri-data to be shared on the system is presented below in figure 4. A set of domain entities that will be used to represent the types of datasets in the system are presented in the conceptual data model along with the logical relationships they have.

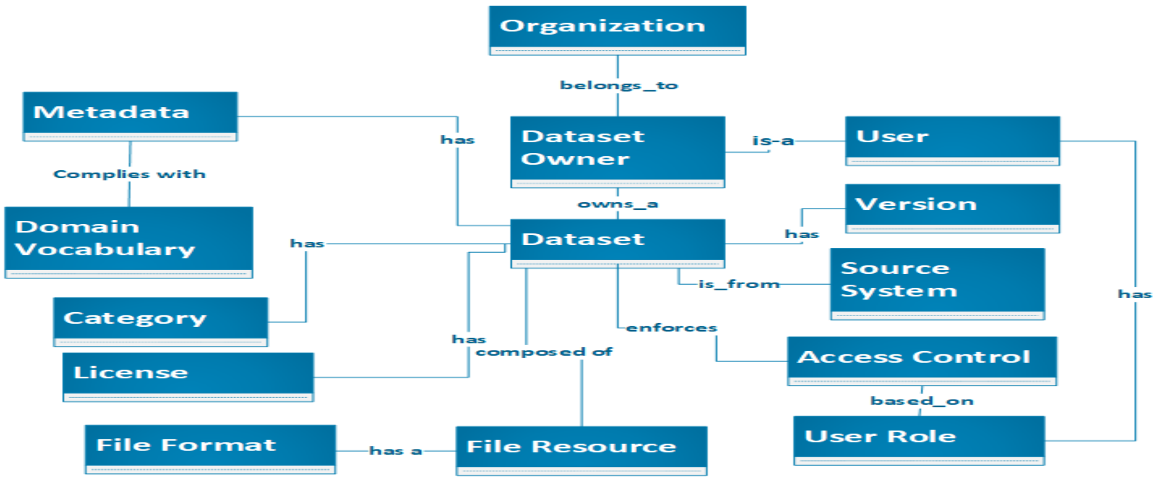

Figure 4: Conceptual data model

As can be seen in figure 4, collaborating organizations in the national Ag-datahub can have dataset owners who will share the datasets in the system on behalf of the organizations. Such owners are expected to have a user account that will be used for authentication and authorization purposes. User accounts are means to identify dataset owners and dataset consumers. User accounts, depending on their owners, are associated with one or more user roles that have predefined access permissions in the system. Users will be able to access services they are allowed to in the national Ag-datahub GUI or machine API through authorization processes.

A dataset is a pool of one or more file resources coming from one or more source systems with a known format/structure. When data is shared on the national Ag-datahub, a metadata profile will be created based on a domain specific set of vocabularies for the agriculture sector. Compliance with standards will be ensured with the development of a national standard for data of the agricultural sector in the country. This ensures conformance to data quality standards and interoperability requirements expected in the system. Moreover, every dataset owner determines the license type for their datasets so that usage restrictions of the data sources are properly enforced.

Table 3 below provides a catalog of the concepts that are presented in the conceptual data model of the Ag-datahub shown in Figure 4.

Table 5: Description of Elements of the Conceptual Data Model

Class	Description	Related to
Dataset	Entity that represents the logical grouping of file resources to be shared in the national Ag-datahub.	Dataset Owner, Resource File, versions, Access Control, Metadata, License,
Version	Logical entity that represents the historical versions of a dataset in the national Ag-datahub	Dataset
Dataset Owner	Entity that represents individuals who share Agri datasets. The individuals may share the datasets on behalf of organizations, or they may share them as individual contributors.	Dataset
Organization	This is an entity used to represent government and non-government organizations in Ethiopia working in the agriculture sector, and that wish to share their data in the national Ag-datahub. Dataset owners may be affiliated to any of these organizations.	Dataset owner
Category	This is an entity used to represent the specific subject within the Agri sector a particular dataset falls under. Dataset can fall under one or more categories.	Dataset
Metadata	This is an entity used to represent the overarching metadata profile with which a dataset is shared.	Dataset
Domain Vocabulary	This is an entity used to define a set of standard vocabularies for a dataset category(subject).	Metadata
License	This entity represents the licensing type set to a dataset	Dataset
File Resource	This entity is used to represent the individual file logically grouped under a particular dataset.	Dataset
File Format	This is used to maintain the type of data format or structure found in a file resource.	File Resource
User	This is an entity used to represent human or software agents	User Role, Dataset Owner
User Role	This is an entity used to represent the predefined set of user types that will have specific access privileges in the system.	User
Access Control	This entity is used to represent the specific permissions for a particular user role. Users will have access to the system services based on these permissions.	User Roles
Source system	This entity represents the address/name of the systems from which a dataset is pulled into the system.	Dataset

2.1.10. USER INTERFACES

The Ag-datahub will have various UI that are organized based on the purpose they provide. The type of UI presented for the users is determined by the user's access role. All/Anonymous users are presented with the home page. Login is not a must to land on the 'Home' page. Tasks such as System admin, configuration for data extraction/ harvesting, transformation, harmonization, data quality verification, sharing local data, and accessing data for DSS require login and authentication of the required service. Widgets for social media, feedback, voting datasets, dataset request, contact, etc. should be consistent in pages applicable.

All UI pages should follow a consistent visual theme design and should be based on the predefined template. Pop-ups displaying notifications, error messages, or warnings should use the same theme and template standard as the rest of the UI in the application.

Users testing will be conducted to ensure the user interface design is clear (simple, intuitive to use, commonly understood vocabulary), complete (users can perform all functions from the interface), and consistent (color, style, layout, menus, links, buttons, and wording are the same throughout the system).

Table 6: User Interface Descriptions

UI	Description	Accessible To
Home	This is the default UI presented for all users visiting the ag-datahub. It should consist of a standard home page menu, button, and links to various UI with the appropriate access privilege. It should provide a login option for registered users, allow new user registration, and manage user profiles.	All users of the ag-datahub
Manage User Account	This is a dashboard for administrative tasks. It should provide the ability to manage ag-datahub users, manage roles, manage groups, manage permission, manage access, manage group assignment,	System Admin
Datasets	This UI should be based on the predefined template for the nature/formats of the data.	Public User, Sector User, or anonymous-accessed data
Configure API Harvesting	UI is only accessible for Data Manager roles. It should provide fields, a data input mechanism, and drop-down options to select predefined parameters, and set a schedule for data harvesting.	Data Manager
Configure Extraction Pipeline	UI that enables a data manager to configure a data extraction pipeline. It should allow us to set, input and/or select the required parameters for data	Data Manager

	extraction tasks.	
Transform Data	This UI should provide the ability to run transformation jobs on a given data. It should provide an appropriate interface to set and input parameters and run data transformation. It should support various transformation options.	Data Manager
Harmonize Data	UI should provide an appropriate interface to set, input and run data harmonization jobs.	Data Manager
Verify Data Quality	UI is accessible for data curators to perform data quality and verification. It should provide the appropriate interface to set and check data quality for completeness, and accuracy.	Data Curator
Search Data	The UI for searching the ag-datahub repository with search text/term. It should be based on predefined search templates.	Public User, Sector User
Query Data	The UI for running queries may include the ability to query the database using the predefined query templates.	Public User, Sector User
Browse Data	The UI for data browsing should provide the ability to browse ag-datahub repositories using the predefined query templates.	Public User, Sector User
Download Data	UI that should provide the required interface for downloading data.	Public User, Sector User
Explore Data	UI with all the required interfaces to explore data in various options.	Public User, Sector User
Visualize Data	This UI provides an appropriate interface to present the data using visualization tools. It should present the data with a default tool.	Public User, Sector
Access Data		
Report	UI template to display Ag datahub data usage statistics.	All users

2.1.11. SOFTWARE INTERFACES

The National Ag-datahub interface should follow the standard MVC (model-view controller) design pattern for rendering UI and modeling data objects. The software interfaces should provide the ability to connect through a direct database, API or conventional file management system to connect to the source system to extract/harvest relevant data.

2.1.12. COMMUNICATIONS INTERFACES

Communications interfaces refer to the requirements associated with any communications functions required including e-mail, web browser, network server communications protocols, electronic forms, and so on. The Ag-datahub shall apply REST-based API to extract and harvest data on a scheduled basis from various source systems as well as the mini data hubs that will be placed at the key partner institutions.

2.1.13. QUALITY ATTRIBUTES

The quality attributes, also known as the non-functional requirements of the system, that deal with the quality with which the functional requirements of the Ag.-datahub are delivered are described below from the perspectives of security, scalability, availability, reliability, usability, and maintainability.

Security: access to the resources in the system will be properly managed, and as per the access requirements, authentication methods will be used. Furthermore, access by authenticated users will be controlled to make sure users are presented with data based on their access privileges.

Scalability: the system will be designed to avoid degradation of performance with the growth of the data of the system and its expansion.

Availability: near real-time access to the data in the source system will be designed to be achieved through scheduled harvesting and extraction processes.

Reliability: reliable tools and technologies will be selected for the development of the Ag-datahub. The data that will be consolidated at the datahub will also be properly managed to ensure users' trust in the quality of data at the datahub.

Usability: the user interface of the system will be designed with user-friendliness and ease of use into consideration.

Maintainability: the system will be designed with potential future changes and improvements in mind to make the maintenance process much more efficient.

2.2. DESIGN OF SOLUTION ARCHITECTURE AND CONFIGURATION OF THE SYSTEM DEVELOPMENT PLATFORM

2.2.1. SOLUTION ARCHITECTURE DESIGN

The **national ag data hub** as shown in figure-1 above is expected to harvest data from Source Systems (ATA, EIAR, GII, other external sources) and mini-data hubs that will be developed as part of this project. After data acquisition from different systems/sources, data will be transformed, annotated, integrated and stored to provide seamless access to data through different exploration and visualization interfaces. Mini-data hubs will also be developed for some agencies with no system to produce and share agricultural data. The system will have different data access options for machines and human users. In order to implement the system, a high-level solution architecture and detailed design of the components is completed in this phase.

The technological options selected for this project combine features of both public data and enterprise datahubs. Innovative software development framework, i.e., *microservice architecture,* is adopted. This architecture enables the creation of multiple loosely-coupled software programs that operate independently but can be harnessed to work together to address the overall requirements of the national ag- datahub. As a result, the approach that was found to be more applicable to the context and requirements of the Ag. datahub is an amalgam of multiple open solutions that address the multi-faceted aspects of the system in a concerted manner. In this regard, the *microservice architecture* that enables the creation of multiple loosely-coupled software programs that operate independently but can be harnessed to work together to address the overall requirements of complex systems is found to be the most optimal technological solution. The general design architecture of the proposed technological solution, which is based on the microservice architecture, is shown in figure 5.

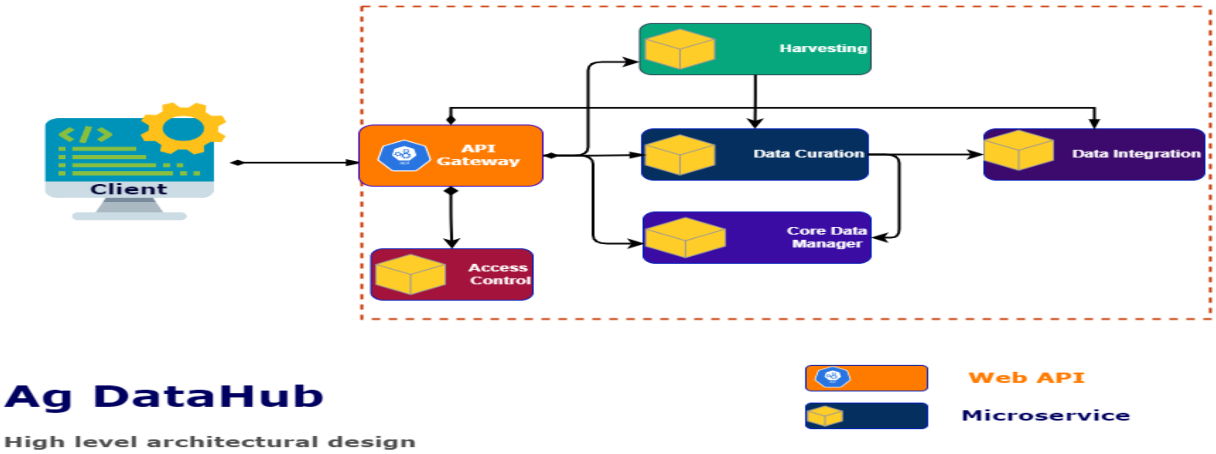

Figure 5:The solution design architecture

2.2.2. DEVELOPMENT SERVER CONFIGURATION

The development server is configured to meet the requirements of different teams. Two servers that will be used redundantly are configured for this project. One of the servers is hosted in the MOA datacenter while **ACATECH TECHNOLOGY PLC** has configured a separate virtual server for development and testing. The development server configuration has two main components: user configuration and system configuration. The User Configuration in this case is the process of creating a separate user account and assigning respective access privilege for different team members of the project while the system Configuration is the process of allocating computational resources since the different services require different requirements. Thus, based on each team requirements, the two servers are configured with

- Different database technologies (ArangoDB, Postgresql)
- Version Control System (Git)
- Open-source core data management framework
- A distributed event streaming platform for inter-service communication(kafka)
- A software containerization platform(docker)

2.2.3. CREATION OF TEAM COLLABORATION/PROJECT PROGRESS TRACKING TOOLS

Team collaboration and tracking tools have been configured to help us monitor the progress of the ag-data hub development project. as can be seen in the screenshot below, the project has now reached a 20% completion rate. This is computed based on the number of activities that are completed in the last three months from the total project estimated activity and time plan.

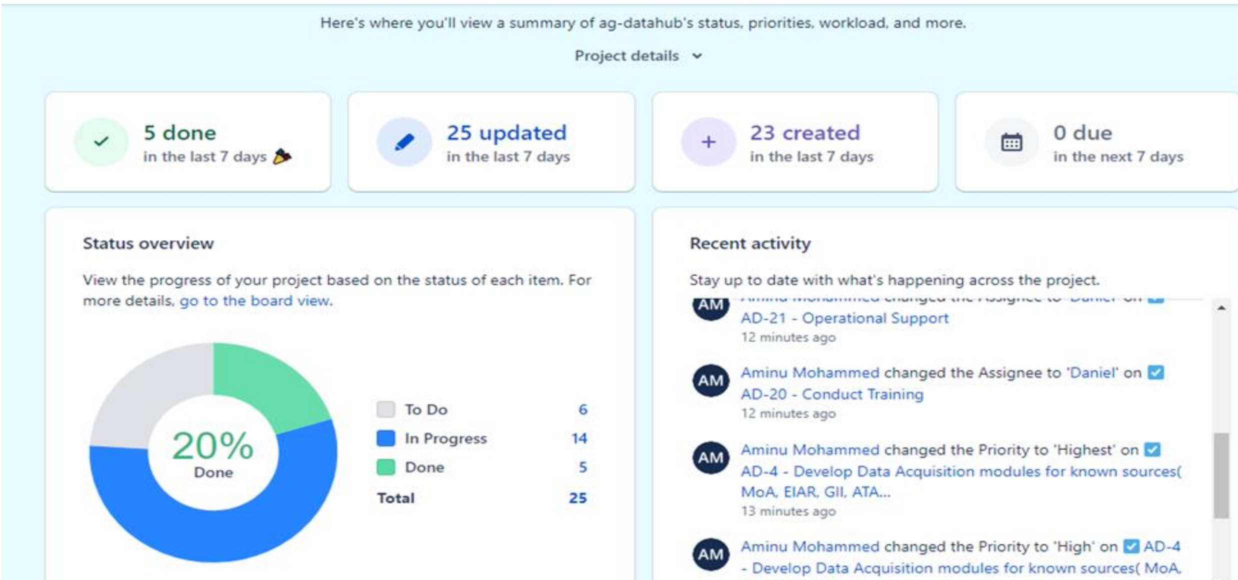

Figure 6: the progress of the ag-data hub development project

The project management team has also prioritized the tasks and assigned prime responsibility for each task as shown below. This makes it easy for us to get an overview of the entire project status.

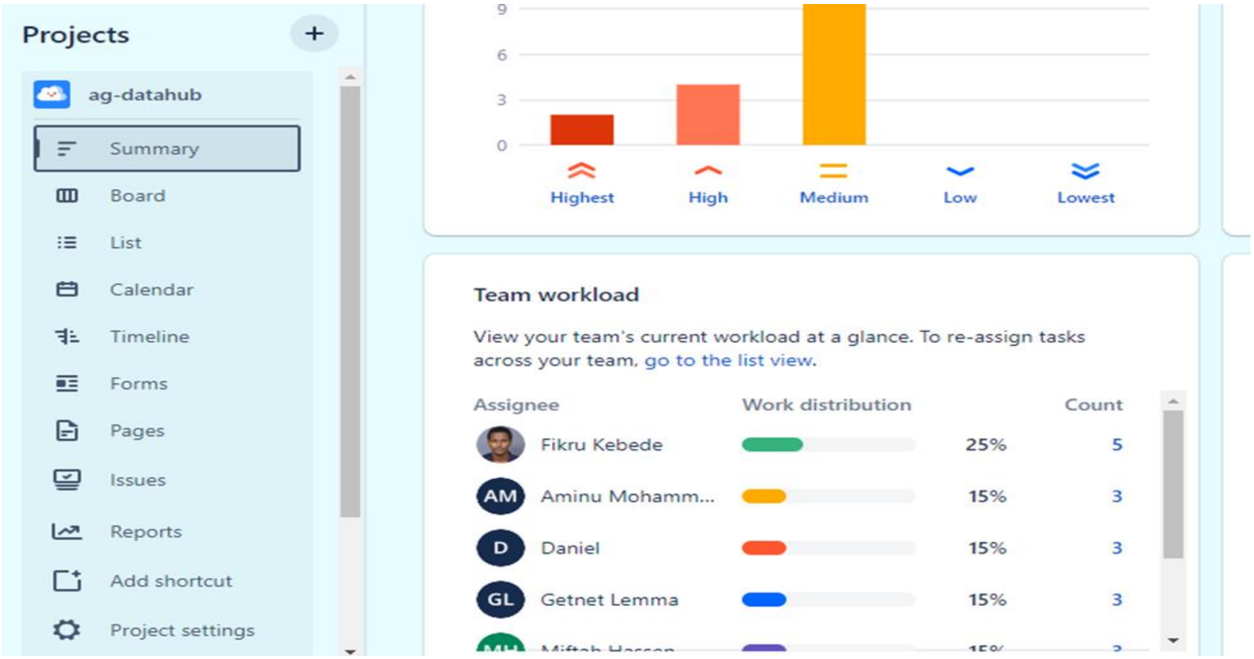

Figure 7:Team Workload

As shown in the screenshot below, of the 25 grand tasks in the ag datahub development project, 5 have been completed successfully, 14 are in progress (some activities are done phase by phase), and 6 activities are yet to be started in subsequent phases.

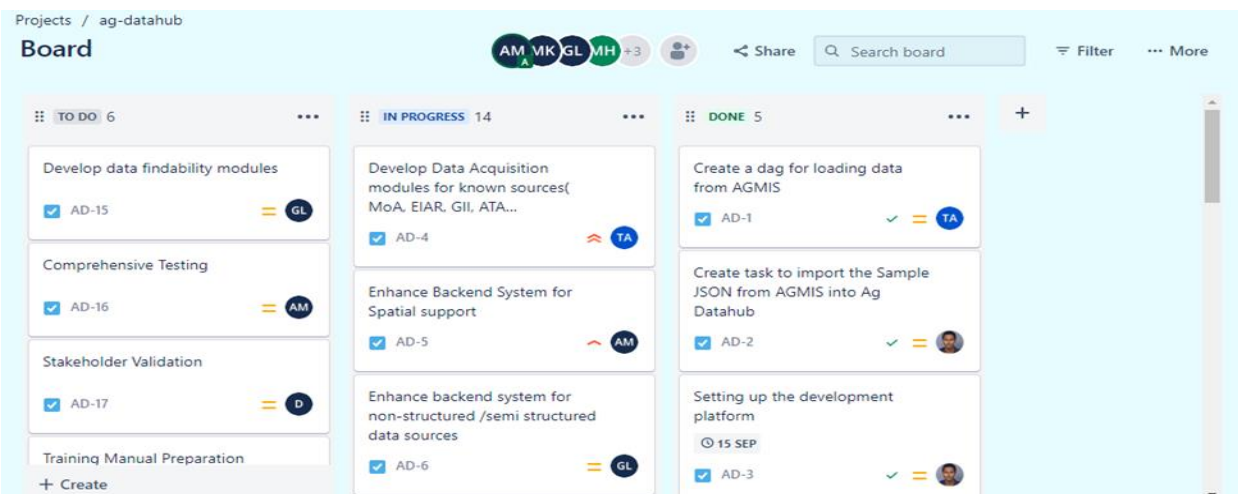

Figure 8: Tasks in the ag datahub development project

Meanwhile, we have also created source code control repositories both on GitHub and bitbucket and Team Communication on slack.

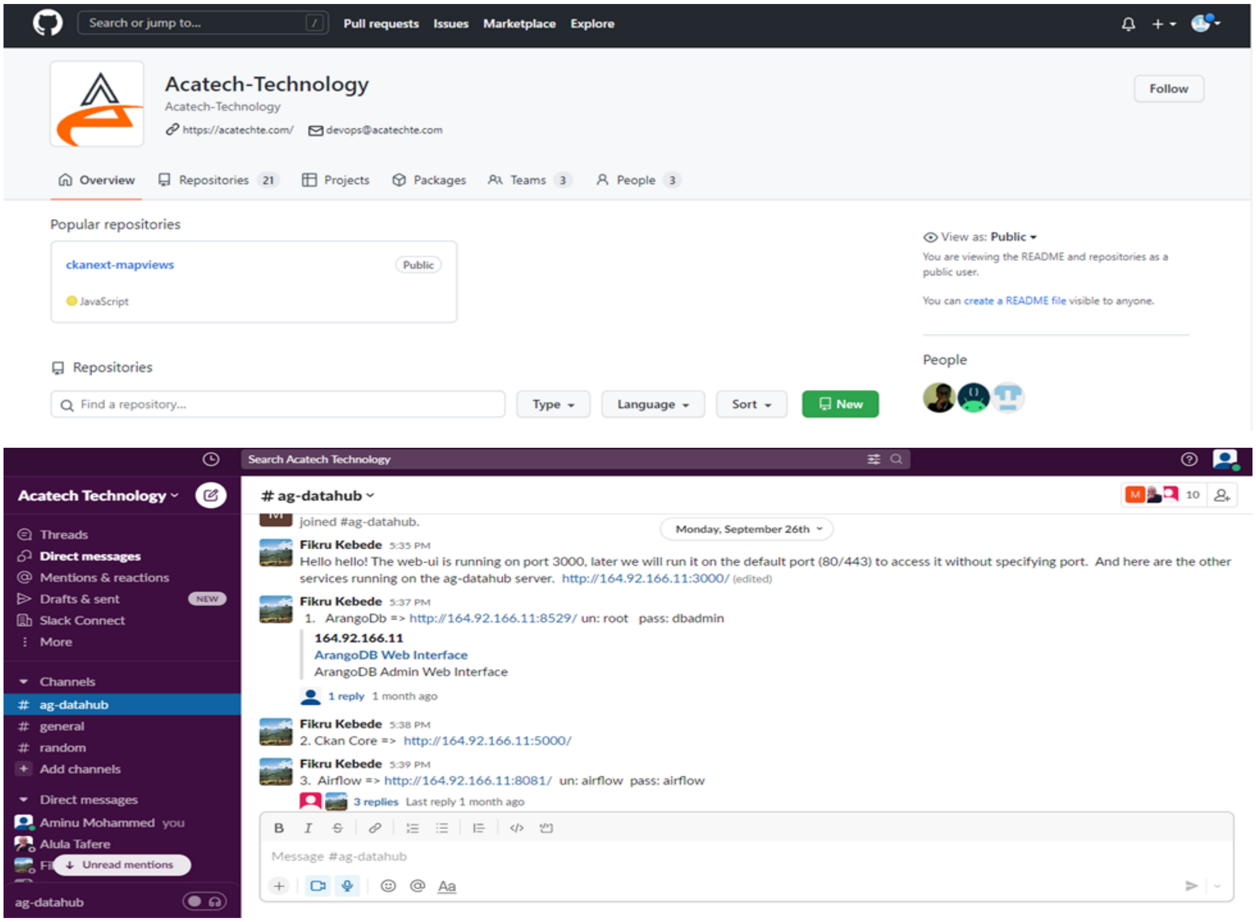

Figure 9: Source code control repositories and slack

2.3. DEVELOP/ADOPT METADATA STANDARD AND INFORMATION ARCHITECTURE

The main purpose of this section is to develop three kinds of deliverables.

1. The first involves designing an information architecture whereby data relevant to Ethiopian agriculture is presented in the main ag datahub interface.
2. The second is designing the Ethiopian agriculture data standard for the data integration service
3. The third is adapting a metadata standard for the core data management service

However, it is not the aim of this phase to come up with a completed product for the above three deliverables. It is rather aimed at an initial idea and implementation plan of the three products in the subsequent part of the project.

2.3.1. INFORMATION ARCHITECTURE

Therefore, an initial Information architecture that will be refined throughout the entire project life cycle is developed. The Ag datahub will contain a wide range of datasets and other resources in various formats. The datasets are grouped into five main categories. The major data cataloging themes identified include Natural Resource, Agricultural Input and Output Marketing, Agriculture Development, Livestock Resource and Cross-sectional. Each of the main categories are also further divided into more themes. Table-x and Figure-x shows some of the sub themes and implementation of the architecture in the national prototype respectively

Table 7: Information Architecture of the Ag datahub

	Theme	Subthemes
1.	Natural Resource and Food Security	Agricultural Land
		Water Resources (Lakes, rivers, swamps, ponds,)
		Forest
		Conservation
		Watershed
		Irrigation
2.	Agricultural input and Output Marketing	Fertilizer
		Seed
		Pesticide
		Mechanization
3.	Agriculture Development	Crop
		Soil
4.	Livestock Resource	Health facilities
		Health professionals
		Disease database
		Livestock population
		Genetic data
		Feed
		Production
5.	Cross-sectional	Infrastructure
		Climate

According to the aforementioned overarching information architecture, the below ag datahub UI has been implemented. As can be seen in the screenshot in figure 10, the main entry points into the system are based on the information architecture defined above.

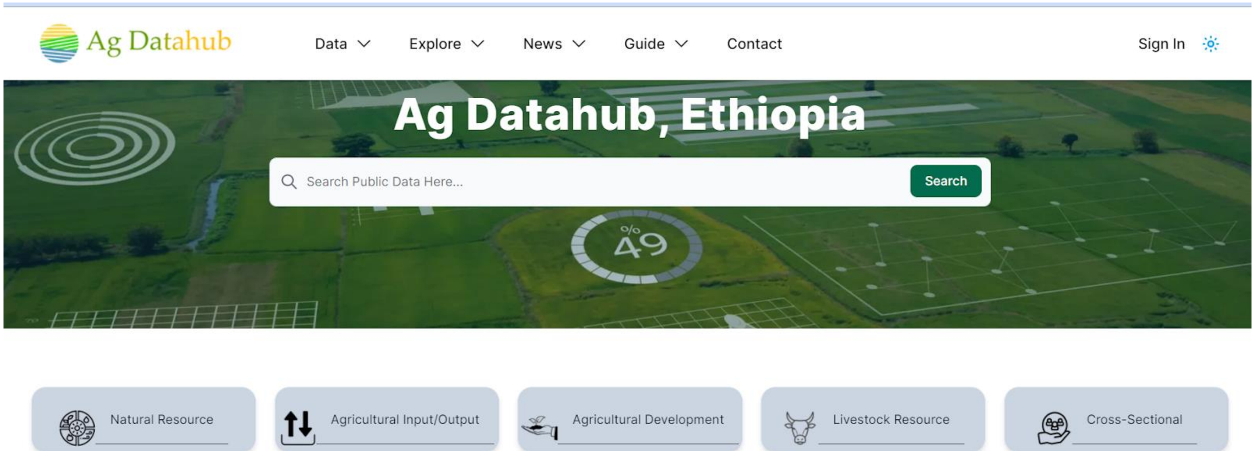

Figure 10: Information Architecture of the Ag datahub

2.3.2. DATA STANDARD FOR THE DATA INTEGRATION SERVICE

The main objective of the data integration service is to create an integrated, multimodal, ontology oriented and query responsive microservice. To make it happen, architecture is designed (figure 11) and the implementation is in progress. Based on the multimodal nature of the data integration service a multi-modal database ArangoDB(figure 12) is selected and configured on the development server.

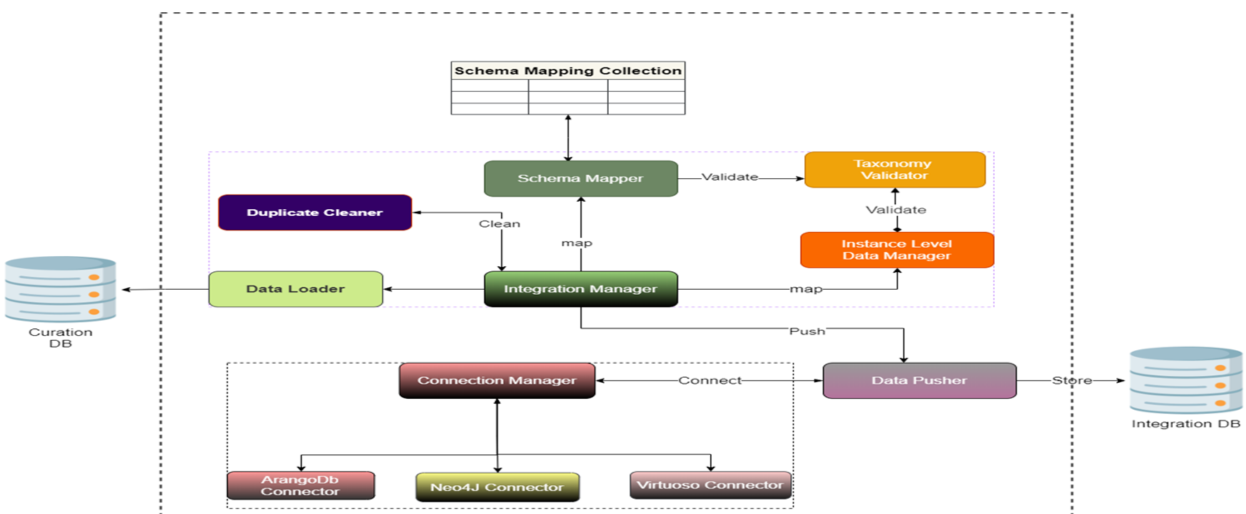

Figure 11: Architectural design of data integration service

After the configuration of a multi-modal Arango DB experimental and production ready datasets are uploaded as shown in figure-13. However, this phase presupposes adapting a data standard that will be used for schema validation in the database.

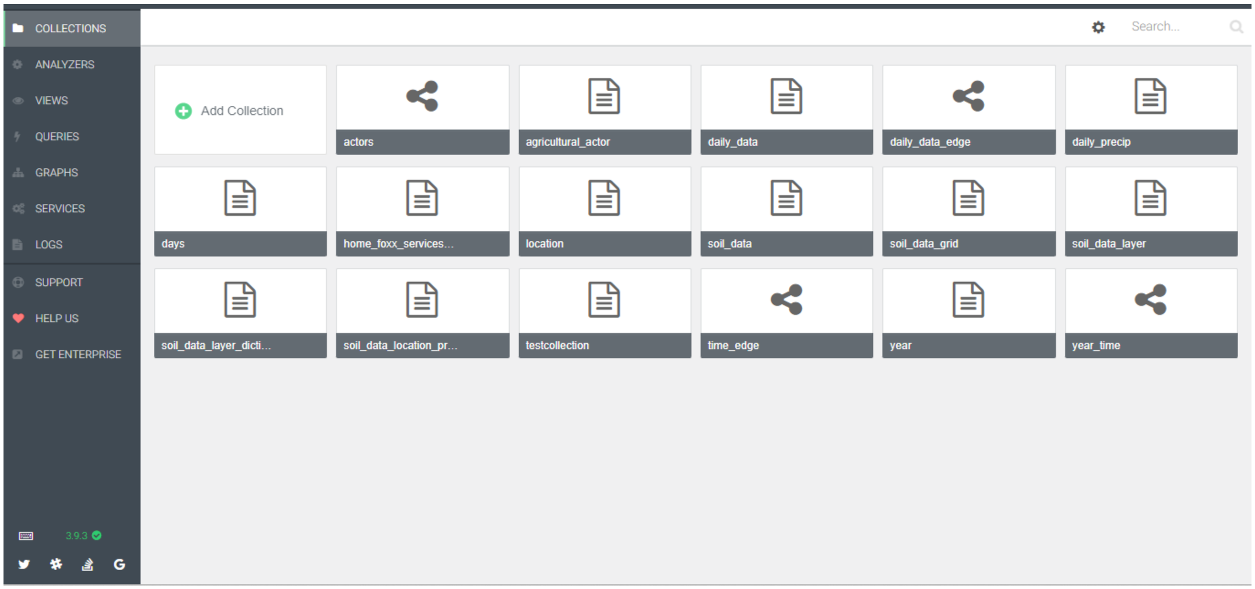

Figure 12:Multimodal Database (ArangoDB)

2.3.3. DATA STANDARDS IN THE AG-DATAHUB

Agriculture data standards have been adapted with the aim of representing the concepts that can be used in the Ethiopian agriculture domain. In the development process, existing standards and common resource description practices currently used in the sector are reviewed and explored. Brainstorming workshops and meetings have been conducted by constituting working groups from different agriculture sub sectors. As a result of this, data standards for the different concepts/sub thematic areas are identified. Please see appendix A for examples of data standards adopted in the ag -datahub.

2.3.4. METADATA STANDARD FOR THE CORE DATA MANAGEMENT SERVICE

One of the microservices in our data hub architecture is the core data management. The prototype of this module is shown in figure-13. In this phase, the purpose is to select and adopt a data object description/metadata standard and use this standard to describe resources. In this project we adopt the Dublin core metadata standard as it is generic enough and contains metadata fields to describe the agricultural data. Some of the metadata elements available in this standard include but are not limited to title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. The system is also designed to allow additional metadata elements. Figure-13 shows the core-data management service and metadata standard will allow in providing standard reusable vocabularies to describe datasets in this service. To be precise data catalog vocabularies as specified in DCAT have been adopted to describe the data in the national Ag-data hub. Implementation is underway to enforce DCAT specification on the core data management module (figure 13) for the national ag- data hub.

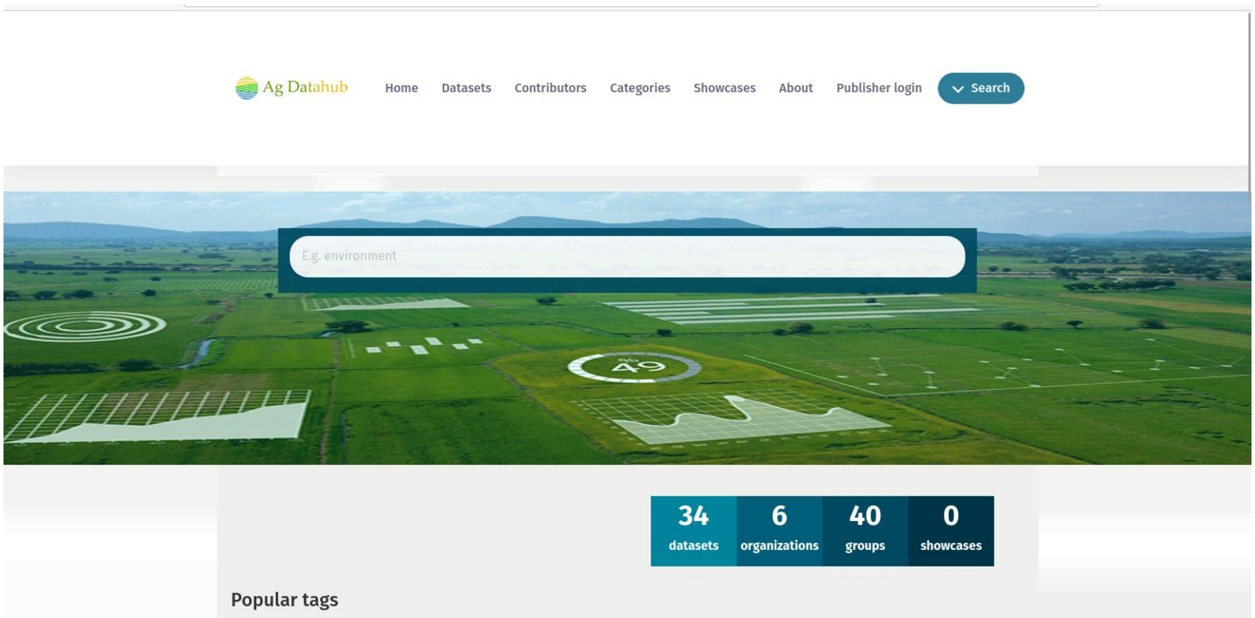

Figure 13:Core-data Management Module

The metadata elements are shown below and additional elements can be extended when the need arise.

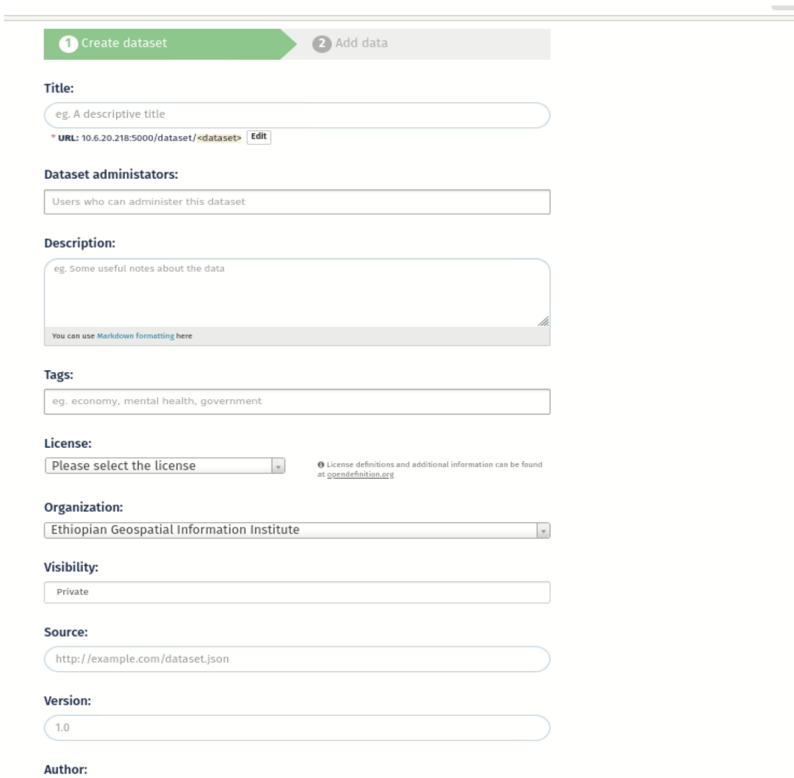

Figure 14:Metadata Elements

2.4. IMPLEMENTING DATA HARVESTING PIPELINES

2.4.1. ARCHITECTURE OF THE HARVESTING MICROSERVICE

The main responsibility of this module is extracting data from the various sources including from external satellite imagery sources relevant for the Ethiopian agriculture, perform the required transformations, and finally load into a staging repository to make it ready for the data curation micro-service that will be done by MOA data managers (MOA data managers will transfer this data to the core data management service and the multimodal data integration module). The architecture of the harvesting module is depicted in the following diagram.

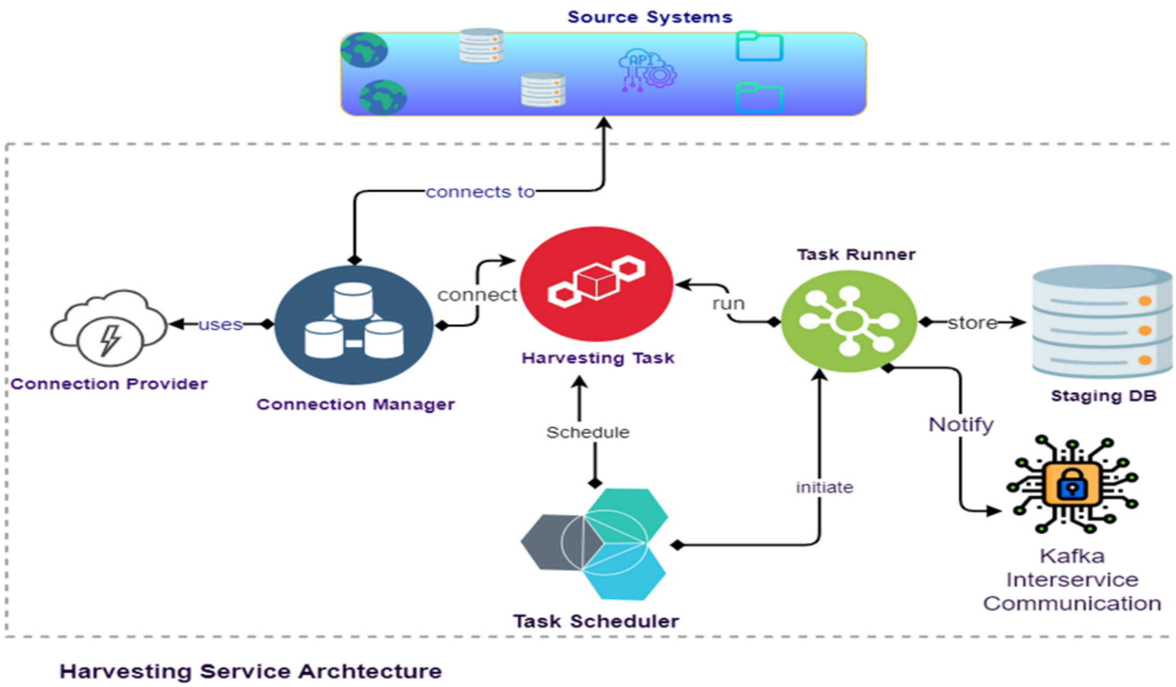

Figure 15:Harvesting Service Architecture

As shown in the figure 15 above, this module extracts data from various data sources. These sources are mainly categorized into - API (link) based, direct database source, web(http) and file management systems.

2.4.2. IMPLEMENTATION OF HARVESTING SERVICE

Based on the nature of the data source system, the extraction pipeline first establishes connections with different database vendors based on credentials received from data providers. Accordingly, in the case of other source systems such as API based, ftp, and http, it uses http and ftp connection type providers to establish connection. The extraction task will not bring everything to the staging database. It rather takes data on views that are from different tables. The screenshot in figure 16 shows the different connections type establishment process implemented in this hub for database, api, ftp, and http source data systems

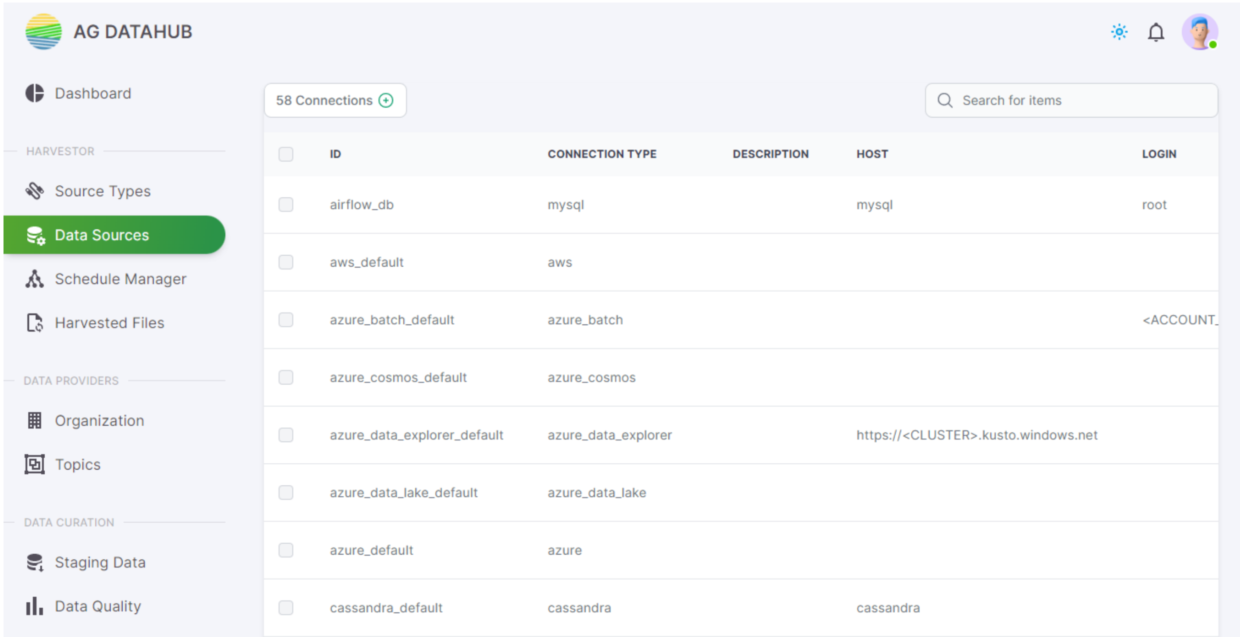

Figure 16: Connection Establishment

As shown in figure 17, a data extraction pipeline is configured and MoA data managers can check the status of this task, whether an error has occurred in the extraction, transformation and loading process and when this task is scheduled to be executed and other actions that a data manager would like to apply.

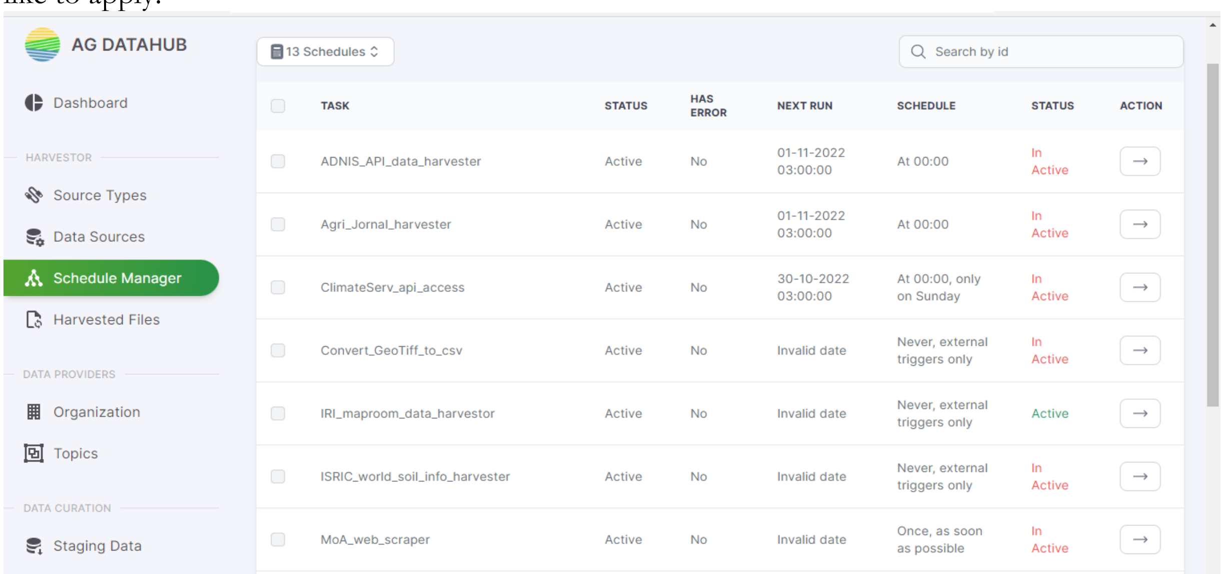

Figure 17: Data Extraction Pipeline Configuration

Once the execution is completed it notifies the next service by saying that a file is uploaded and the next service is ordered to do whatever action it should take as shown in figure-18.

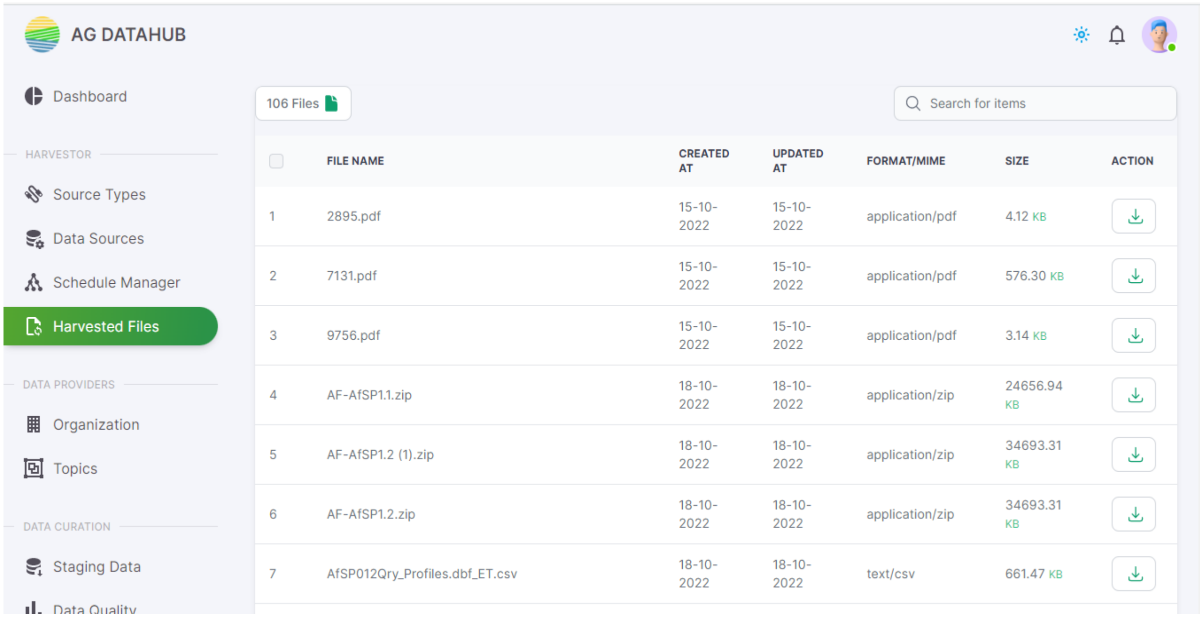

Figure 18: Harvested File

Moreover, harvesting options for sources having APIs is also configured. For instance, the following data harvesting UI (shown in figure 19) is designed to harvest data from mini-datahubs.

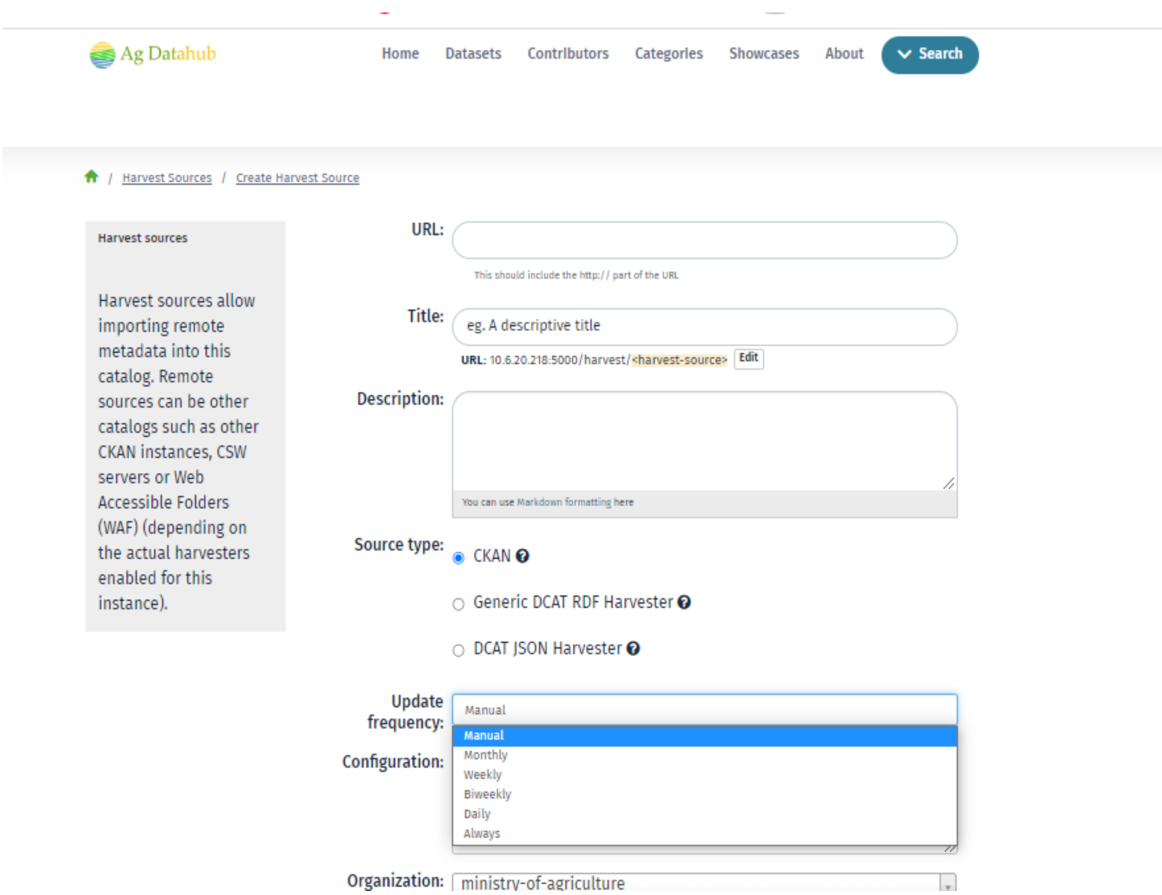

Figure 19: Harvesting Source Configurations

Once a source is configured, its status can be configured as shown in the screenshot in figure 20.

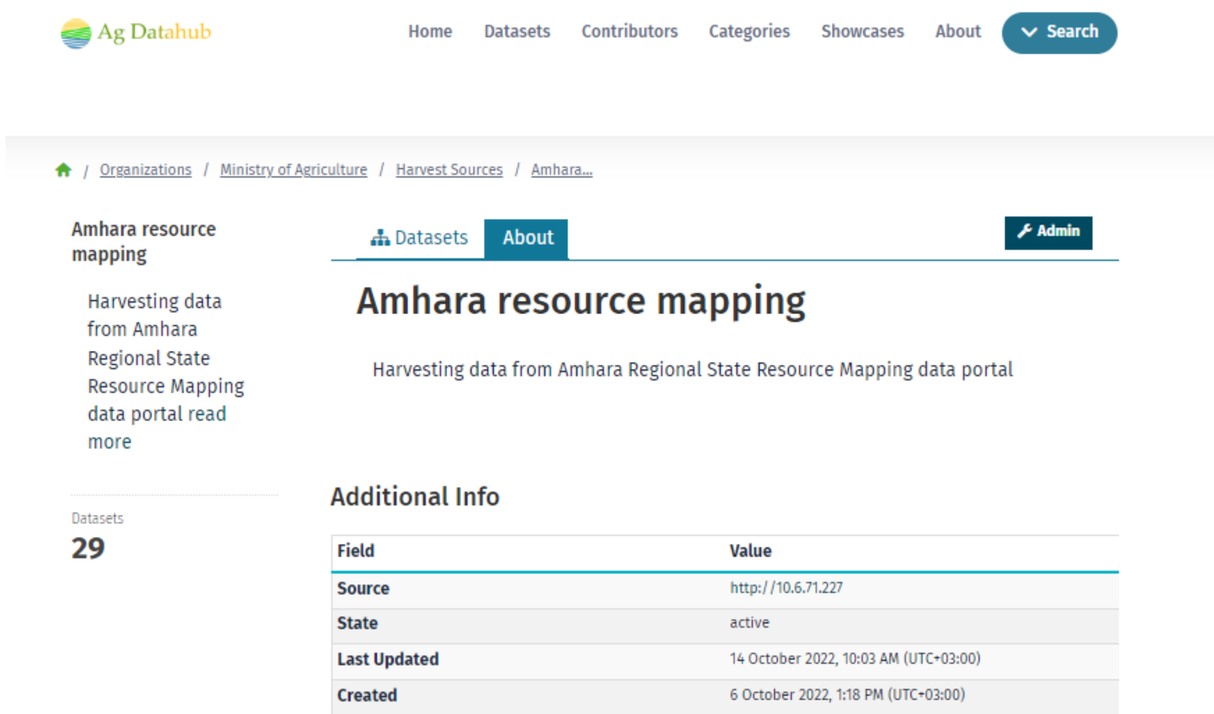

Figure 20: Harvest data from mini-datahubs

2.4.3. BRIEF DESCRIPTION OF HARVESTED DATA

Table 8 shows a summary of data sources where we have established data harvesting pipelines so far.

Table 8: Summary of Data Sources

SOURCE Types	System
MoA source systems	• LMIS -Livestock market Information Systems • AGMIS-Agricultural Management Information Systems • Ethio SIS- Ethiopian Soil Information Systems • Disease Outbreak and Vaccination Reporting system(DOVAR) • Animal Disease Notification and Identification(ADNIS)
Global Climate data sources	IRI, DE-Africa, CHIRPS IRI(http://iridl.ldeo.columbia.edu/maproom),

	CHIRPS(https://www.chc.ucsb.edu/data/),climateSERV (https://climateserv.servirglobal.net/map),
Global Soil Data	ISRIC(https://files.isric.org/public) FAO(https://www.fao.org/soils-portal/en/
Mini-datahub	Amhara Regional datahub
Agriculture Literature	EIAR Institutional Repository,
Commodity and daily currency rate	NB(https://nbebank.com/)
National weather forecast	NMA(http://www.ethiomet.gov.et/forecasts/three_day_forecast)

The format of extracted data is summarized in Table 9.

Table 9:Summary of data formats extracted from sources in table 8

Source systems	data format
LMIS, AGMIS, EthioSIS, DOVAR, ADNIS,	json, csv, table
ISRIC, IRI, DE-Africa, CHIRPS	GeoTIFF, VRT, dbf, shp, net4CD, asc, zip, png
EIAR, DOAJ, Agromet bulletins	pdf, doc
NMA, NB	html,javascript

The harvesting pipeline does not only include extraction of data but also additional transformation tasks in order to make the data useful for data users. Once the data is extracted, two major types of transformation can be done on harvested data: structural and content level transformation. Structural transformation include - e.g., changing rapped/ hierarchical fields into flat fields. Format conversion - e.g. changing GeoTiff to csv, net4CD to csv, shape to csv, asc/ascii to csv, dbf to csv, and accdb to csv. merging/joining - e.g. complete soil data are generated from soil ph, soil profile, soil texture, soil salinity, etc. Resize the data by setting the bounding box of the required area/location - e.g bounding box for Ethiopia is bb = (401141.5244,5322463.1536,1682837.6147,3659193.4181) as shown in figure 21 below

Content level transformation includes anonymizing data value related to sensitive/ private data (such as telephone, disease, species, etc.), convert/standardize measurement unit, fill missing value.

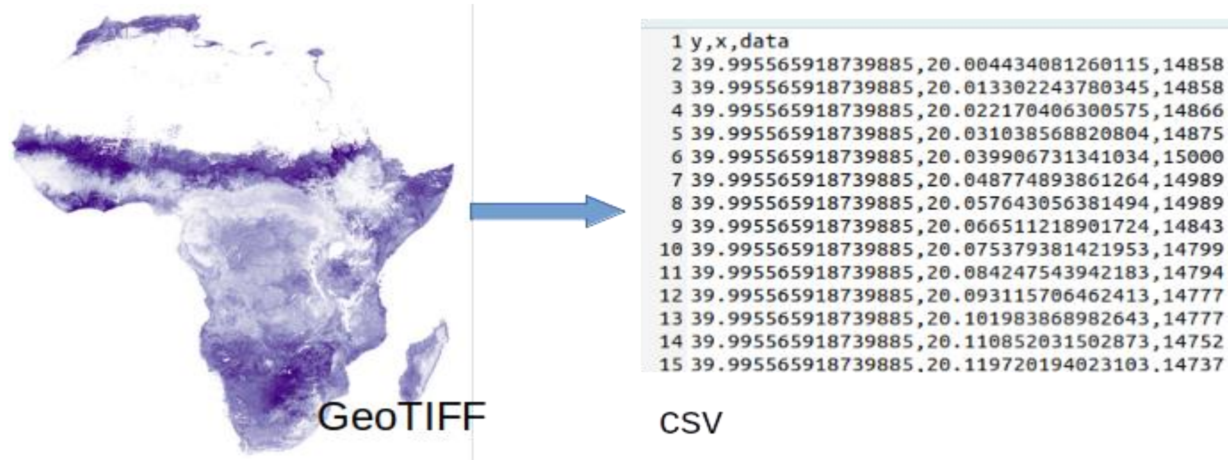

Figure 21: Sample GeoTIFF harvested data converted into csv format

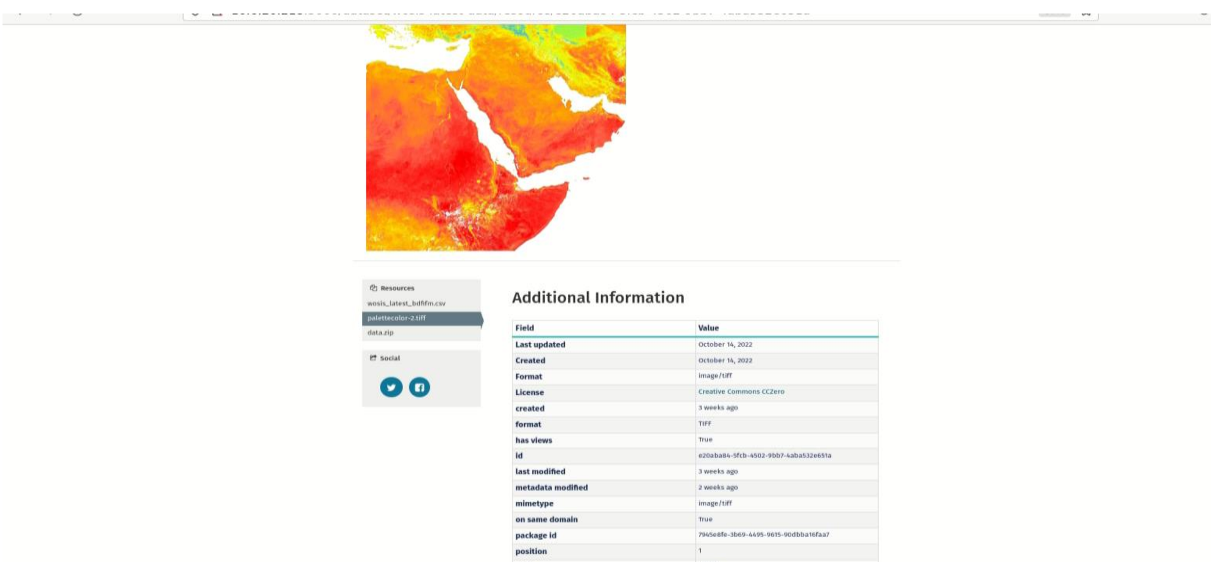

Figure 22: Sample of harvested data in the core data management module

The transformation also includes denormalization of relations from the classical relational database model into a multimodal database as shown in Figure23.

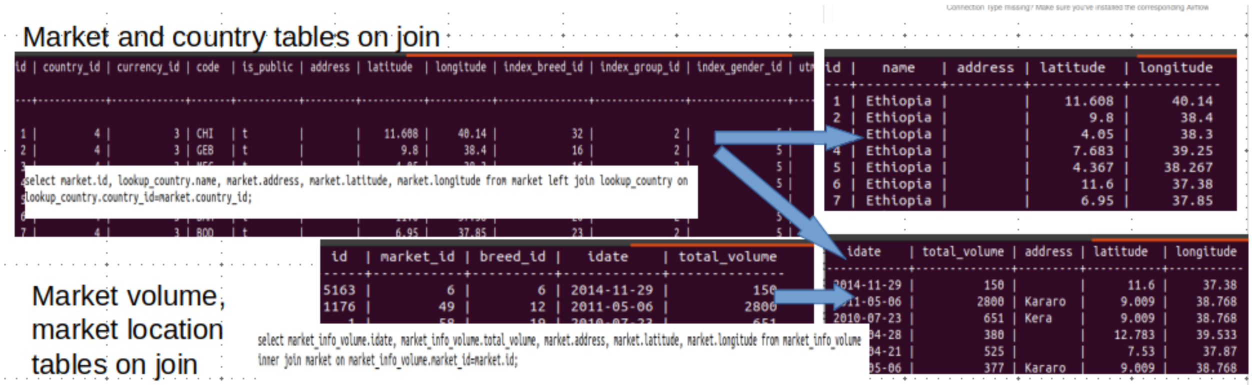

Figure 23: Sample Livestock market data generated by joining and filter multiple tables data

Once the required transformation is made, the data will be loaded to the data curation microservice which will be triggered automatically based on a time provided trigger point or when a human expert triggers the data extraction pipeline. The team has also managed to harvest data from existing mini-datahubs. Such sources have been configured and scheduled in the data harvesting pipeline. As can be seen below, about 29 datasets having more than a thousand files have been harvested from the Amhara regional state data hub with a single runtime.

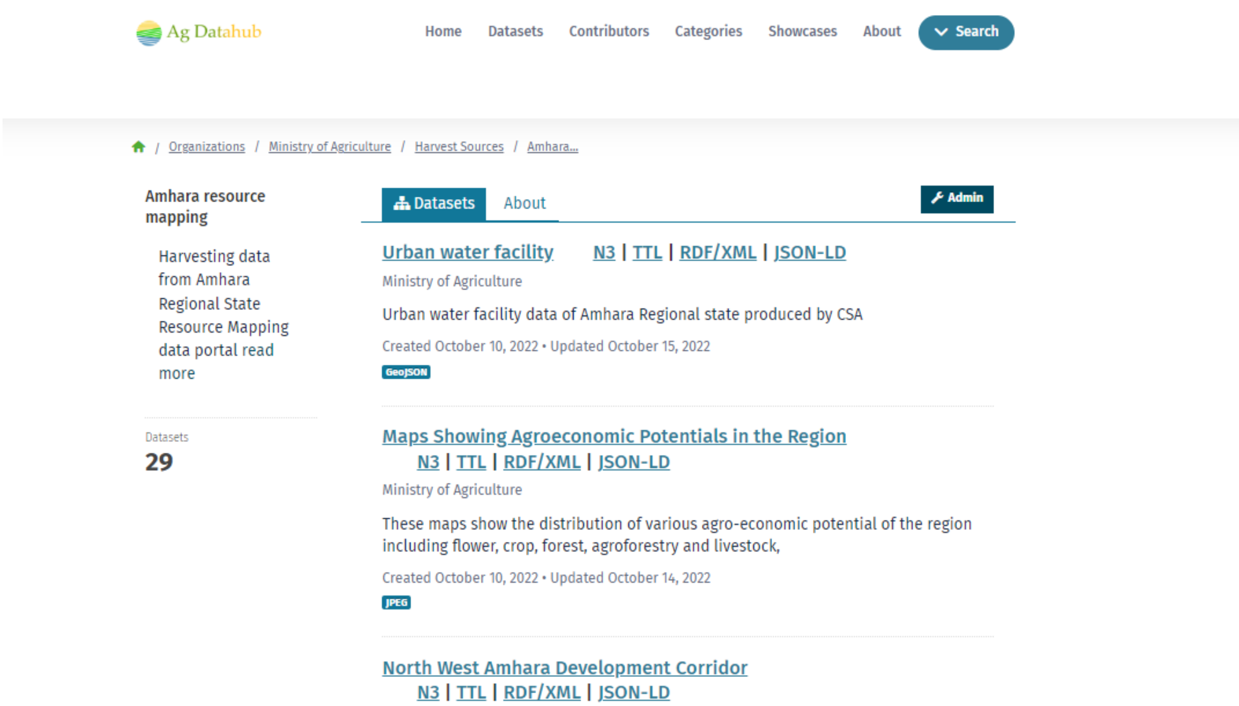

Figure 24: Harvesting Datasets from Amhara Regional State (ARS) Mini data hub

2.5. UI/UX DESIGN (USER INTERFACE AND USER EXPERIENCE DESIGN)

In this section, data access points and interfaces are being designed to make the national ag-datahub accessible for both humans and machines alike. The state-of-the-art UX/UX design patterns and trends are applied to fulfill the various user requirements of the national ag-datahub. Mainly the interface is grouped into two main categories namely web interface and application programming interface (API).

API provides data access mechanism for machines. The primary consumer of the API is the ag-datahub web interface. In addition to that, any kind of application on top of ag-datahub will use the API endpoint. The main structure and some of the API endpoints (such as harvesting service and core data management) are implemented. Web UI: the Web UI is a web based graphical user interface that enables the user to interact with the ag-datahub. There are two main categories of UIs developed in this project.

Admin module: this module is designed for all privileged users such as system administrators, data curators, data contributors, data managers and decision makers in the MOA. The functionality of this module, shown in the screenshot below in figure 25, depends on the role of the privileged user and each user access only for the role specified.

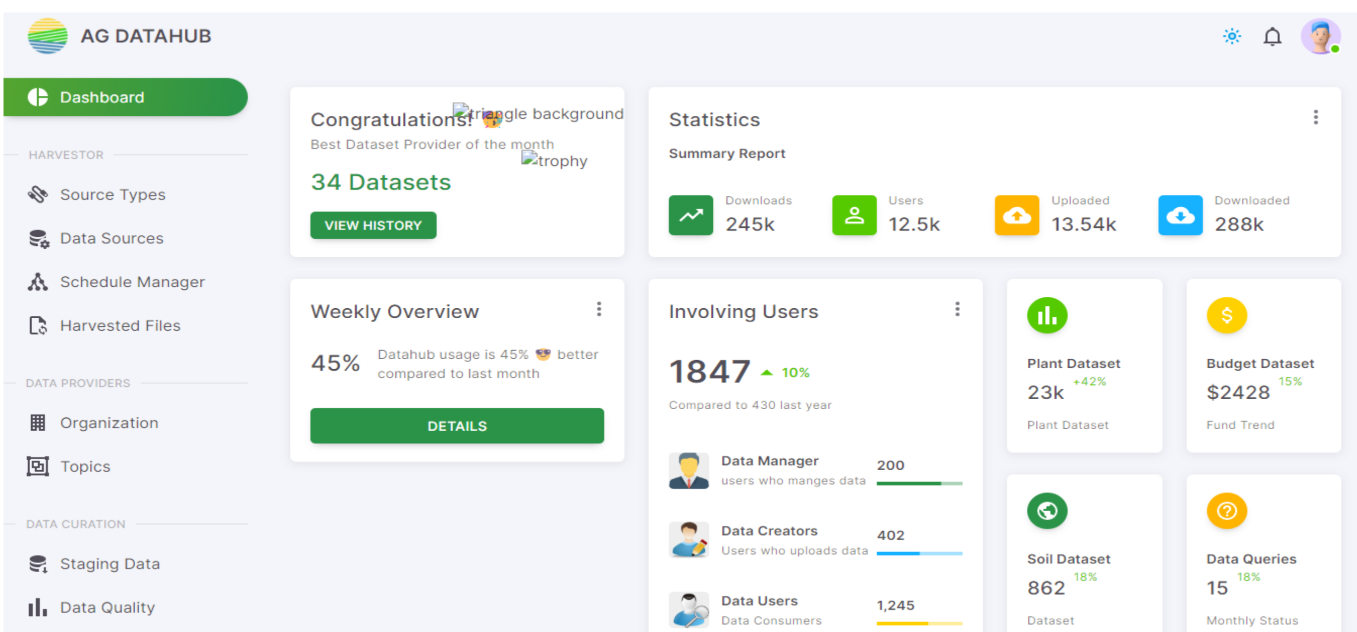

Figure 25: Admin dashboard

Data User Module: This UI module is designed for the general data users. researchers, investors and any interested users can access all public datasets based on the restriction level and license attached to the dataset. The main layout for data visualization, dataset searching and querying interfaces are also defined.

For both the web interface and API, state of the art tools and techniques are applied. The figure below (Figure-26) shows the feel and look of the landing page of the data user web interface of the ag-datahub

Implemented web interfaces so far include the following items in the ist/. A screenshot of the interfaces is attached in subsequent figures.

- Main ag datahub interface / home page,
- Core data management module interface,
- Admin panel dashboard page,
- User authentication and authorization interface and
- Data harvesting interfaces
- Data Curation interface
- Data visualization and reporting interface
- Multimodal query interface

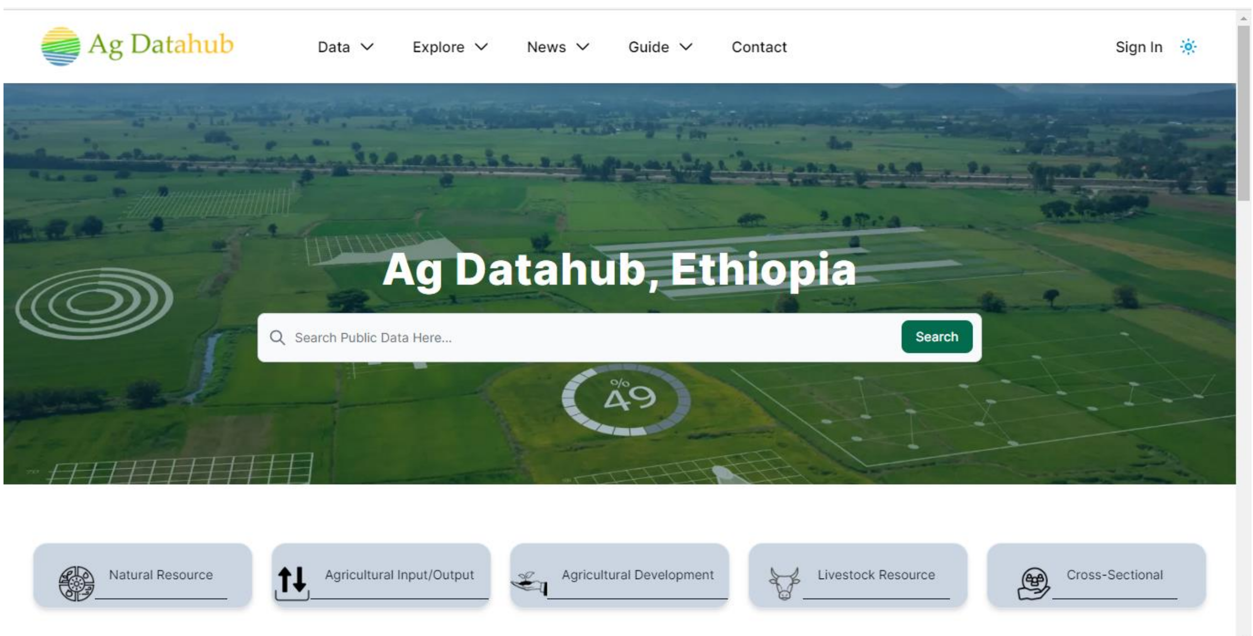

Figure 26: Landing page of ag-datahub web Interface

Clicking on one of the main data categories above, will take the end user to the details page for the selected data category as shown in figure 27.

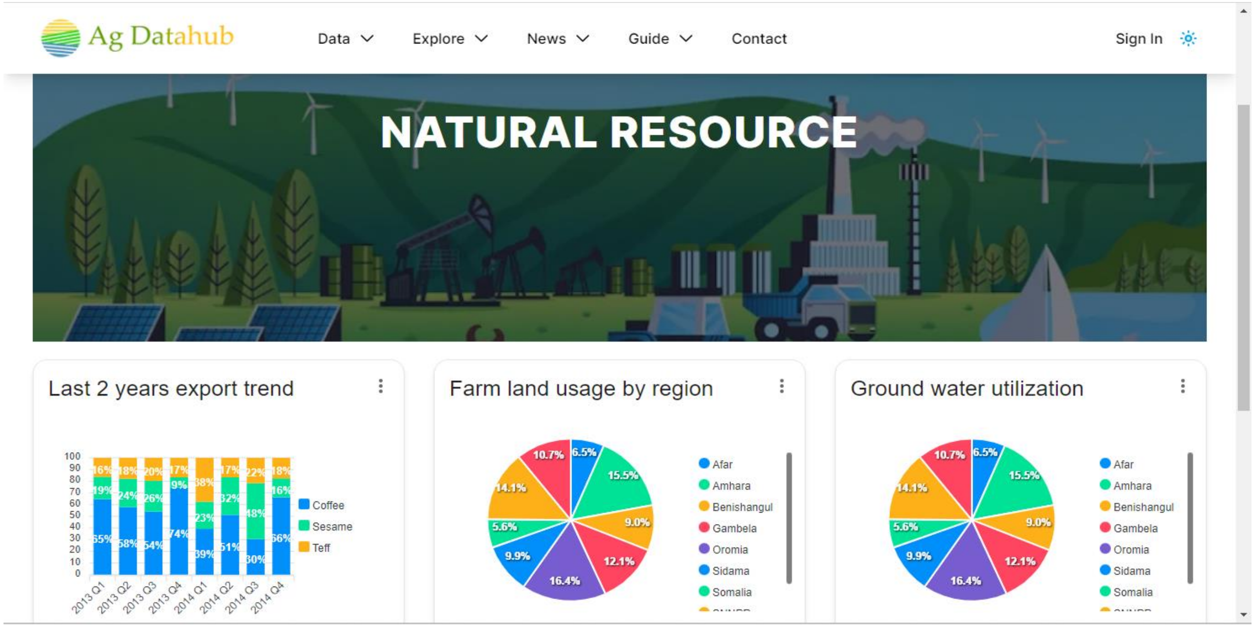

Figure 27: Data Category detail page

Moreover, the UI also provides dataset search and filter options as shown in figures 27 and 28.

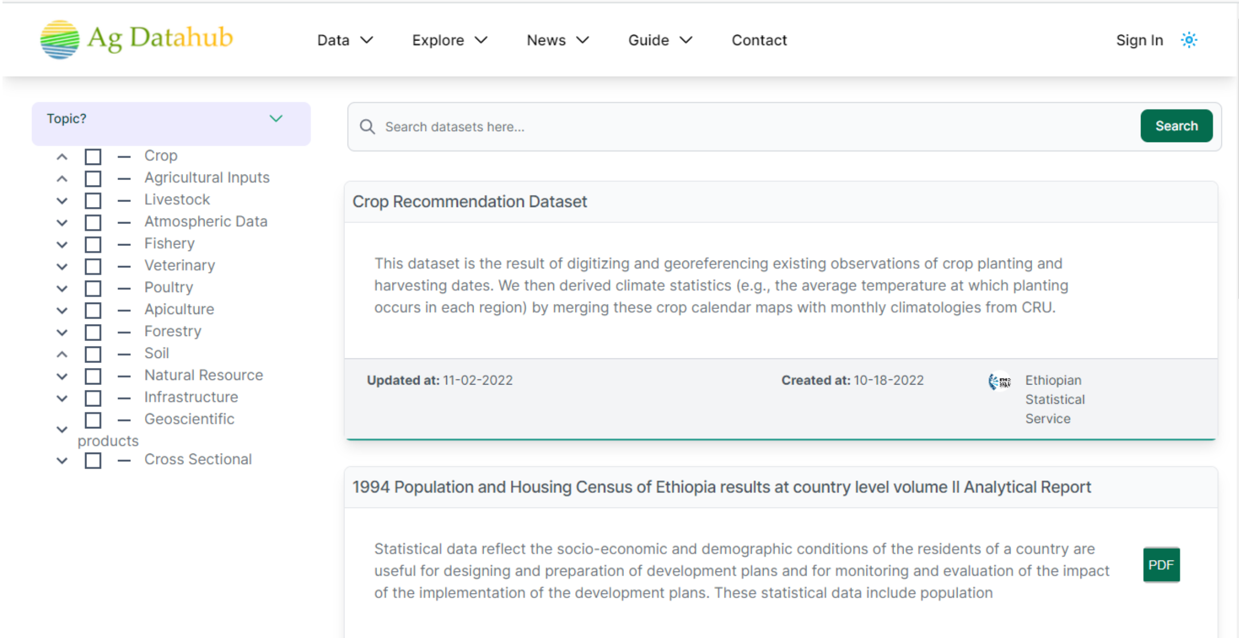

Figure 28: Dataset search and filter page

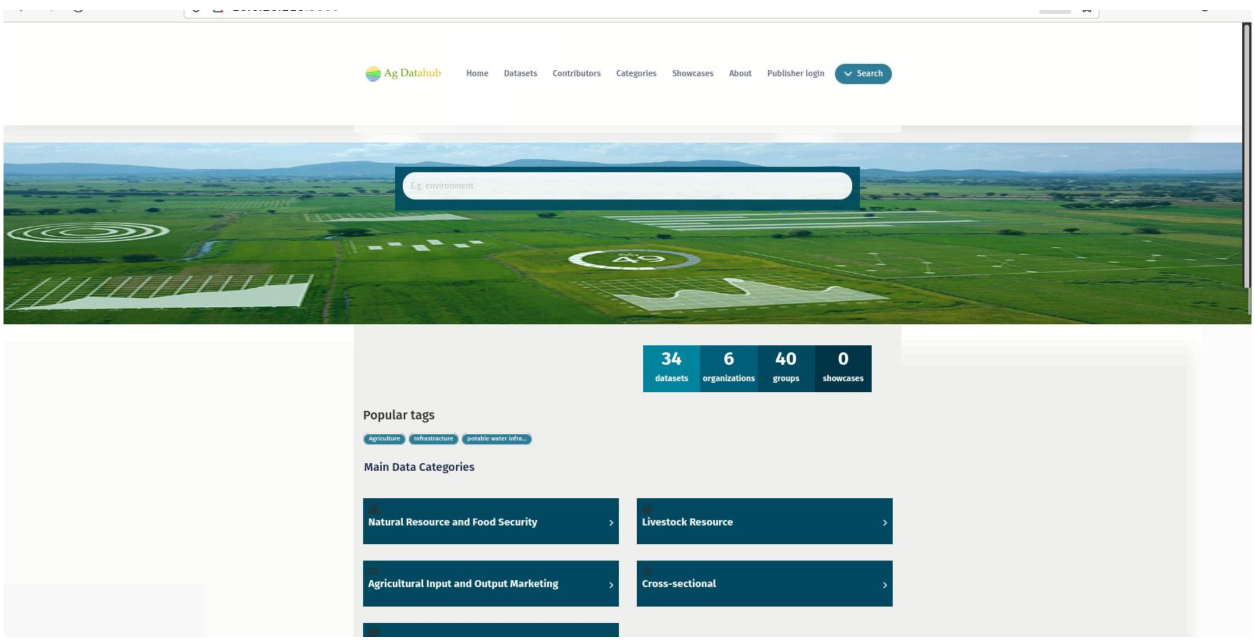

Figure 29: Core data management Microservice

For some advanced users the system also provides a data query interface as shown in figure 30.

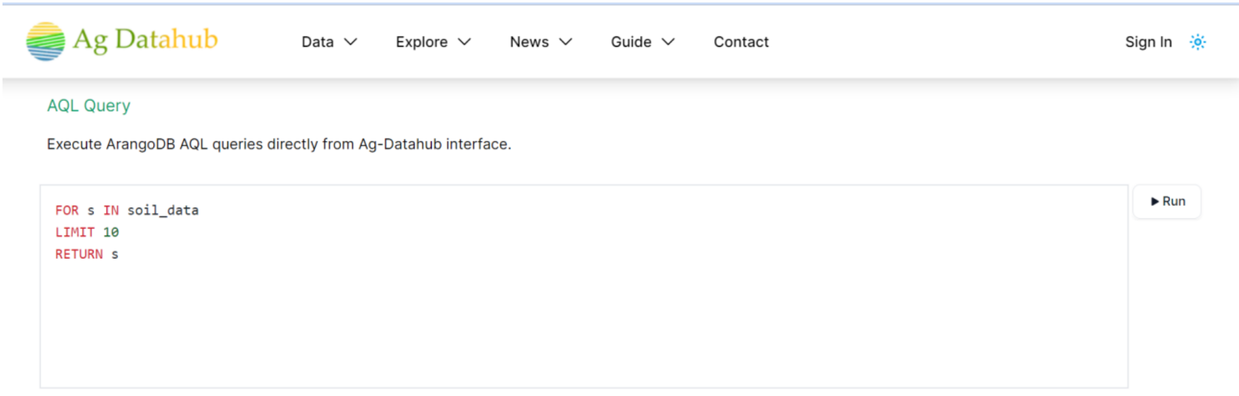

Figure 30: Data Querying Service

What is more, users who wish to access protected data may log into the system through a log on window shown in figure 31. Access will be determined based on rights given to each user.

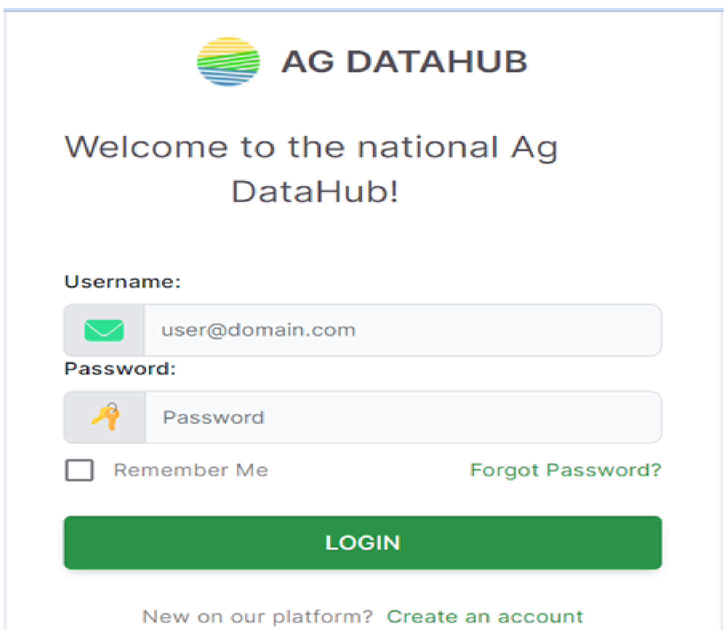

Figure 31: User Registration and Authentication

It is also possible to access the system via an API which is exemplified with the screenshot in figure 31.

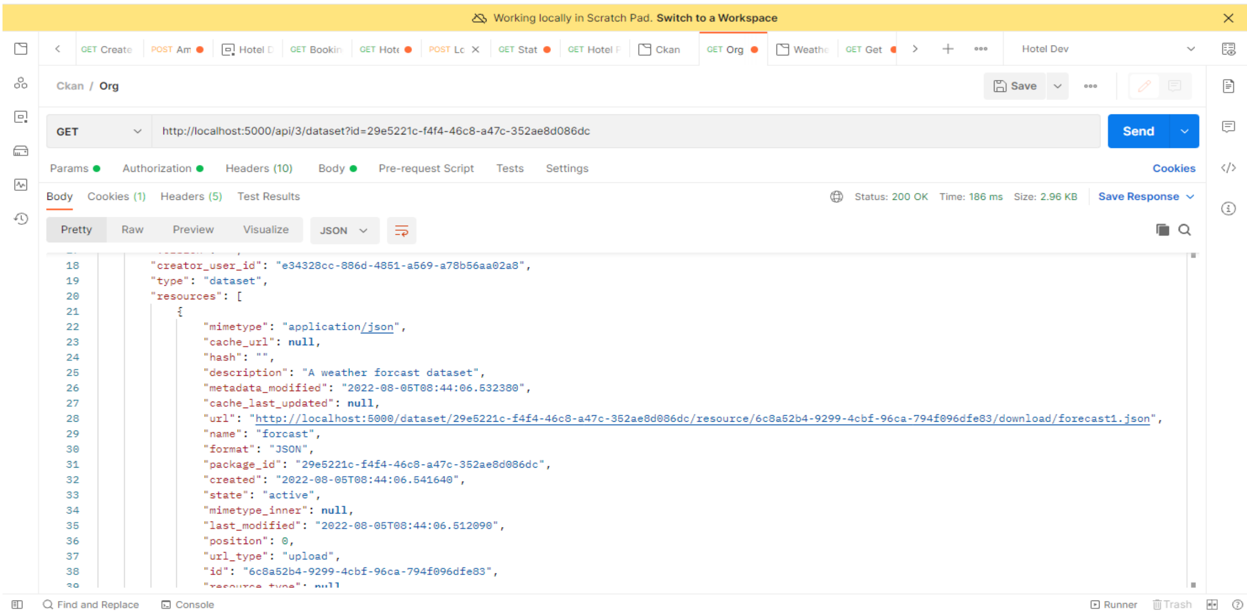

Figure 32: Sample API request

Continuous Integration (CI) AND Continuous Delivery (CD) PIPELINE

To deliver the application frequently, a CI/CD pipeline is implemented. Specifically, the CI/CD pipeline implementation introduces ongoing automation and continuous monitoring throughout the lifecycle of apps, from integration and testing phases to delivery and deployment.

To achieve the CI/CD pipeline plan, the leading DevOps tools are installed and configured on the ag-datahub development server. The table below provides summarized information about the DevOps tools we are using for automation purposes.

Table 10: Summary of DevOps tools used in the ag-data hub development project

SN.	Tool	Description	Status
1	Git	A version control system	Configured
2	Github	Source code repository	Configured
3	Atlassian Bitbucket	Source code repository	Configured
4	Atlassian Jira	Issue tracking	Configured
	Atlassian confluence	Technical documentation	Configured
5	Docker	Application containerization	Configured
6	Jenkins	Integration and build automation tool	Configured
7	Prometheus	Service monitoring solution.	Installed

Currently the CI/CD pipeline is configured for UI/UX (Web interface), harvesting service, core data management service, and api-gateway service.

Interservice Communication

Due to the complexity of the project, the system is decomposed into different subsystems (microservices) based on the functionality/feature. This makes the system more reliable, scalable, and maintainable. Each microservice is a standalone system that has to be deployed separately and there must be a way to interconnect all the microservices together.

To fulfill this requirement, a distributed event streaming platform called Apache Kafka is installed and well configured.

API-Gateway implementation

One of the features of the Ag-datahub project is its scalability. Based on the requirements new services can be added and scaled up or services can be pulled out and scaled down easily without affecting the other services. To achieve this all the services, need to be registered and accessed from a central service registry (API-gateway).

So, all the services are configured with the api-gateway and all microservice endpoints are accessible through that and both the web interface and API requests are configured working fine. In addition to the service registry, the access control service can be performed from a single window. The architectural diagram below shows how the service registry and access control service is configured.

CONCLUSION

This report is a brief description of activities that were performed in the last three months. The requirement specification document refinement, detailed solution architecture, technology choices, configuration of the development/test servers, databases, Installation and configuration of the version control system, source code repository, issue tracking, application containerization, integration and build automation tool, inter service communications are fully completed. The first phases of metadata and data standards adaptation and implementation of these standards within the core data management service and the multimodal database, data hub information architecture and implementation of this architecture in the ag datahub system, data harvesting pipeline implementation, user interfaces/experiences implementation are completed but these phases need further work in the next phase of the project.

APPENDIX:

A: Adopted Data Standards

A:1. Soil Data Standards

Data standards when capturing soil point/grid data is shown below.

Attributes	Code	Data Type	Units
Profile ID	ProfileID	Text	
Longitude	X_LonDD_WGS84	Num double	DecDeg
Latitude	Y_LatDD_WGS84	Num double	DecDeg
Layer number	LayerNr	Num integer	
UpDepth	UpDpth	Num integer	cm
LowDepth	LowDpth	Num integer	cm
Gravel content	CrsFrLabPc	Num double	v%
Sand	Sand	Num double	%wt
Silt	Silt	Num double	%wt
Clay	Clay	Num double	%wt
Texture classes	LabTxtr	Text	%wt
Bulk density	BlkDens	Num double	g/cm3
pH H2O	PHH2O	Num double	
pH KCl	PHKCl	Num double	
Electric conductivity	EC	Num double	dS/m
Electric conductivity saturated paste	Ece	Num double	dS/m
Slbna(meq/l)	Slbna	Num double	meq/l

SlbK (meq/l)	SlbK	Num double	meq/l
SlbCa (meq/l)	SlbCa	Num double	meq/l
SlbMg (meq/l)	SlbMg	Num double	meq/l
$SlbCO_3^{-2}$ (meq/l)	$SlbCO_3$	Num double	meq/l
$SlbHCO_3^{-1}$ (meq/l)	$SlbHCO_3$	Num double	meq/l
$SlbCl^{-}$ (meq/l)	$SlbCl^{-}$	Num double	meq/l
$SlbSO_4^{-2}$ (meq/l)	$SlbSO_4$	Num double	meq/l
Exchangeable calcium	ExCa	Num double	cmolc/kg
Exchangeable magnesium	ExMg	Num double	cmolc/kg
Exchangeable sodium	ExNa	Num double	cmolc/kg
Exchangeable potassium	ExK	Num double	cmolc/kg
Exchangeable hydrogen	ExH	Num double	cmolc/kg
Exchangeable aluminium	ExAl	Num double	cmolc/kg
Exchangeable acidity	ExAcid	Num double	cmolc/kg
Effective CEC	Ecec	Num double	cmolc/kg
Cation exchange capacity (CEC)	CecSoil	Num double	cmolc/kg
Gypsum	CaSO4	Num double	% wt
Calcium carbonate	CaCO3	Num double	% wt
Organic carbon	OrgC	Num double	% wt
Total nitrogen	TotalN	Num double	% wt
Available P	AvailP	Num double	mg/kg
Extractable Fe	ExtrFe	Num double	mg/kg
Extractable Al	ExtrAl	Num double	mg/kg
Extractable Mn	ExtrMn	Num double	mg/kg
Extractable Zn	ExtrZn	Num double	mg/kg

Extractable Cu	ExtrCu	Num double	mg/kg
Extractable B	ExtrB	Num double	mg/kg

A:2. Livestock Resources

The Ag datahub adopted the livestock data exchange standards and specifications published by the International Committee for Animal Recording (ICAR). The aim of this standard is to facilitate smooth exchange of livestock data between different systems, thereby enabling data analysis and manipulation that was not possible with data in silos.

Animal Population/ Registration

Attributes	Data Types	Values	Descriptions
Identifier			
Alternative Identifiers	array of icarAnimalIdentifierType		
Specie	iscAnimalSpecieType	Valid values ("Buffalo", "Cattle", "Goat", "Horse", Pig", "Sheep")	
Gender	iscAnimalGenderType	("Female", "Female.Neuter", "Male", "Male.Cryptorchid", "Male.Neuter")	
birthdate	icarDateTimeType		ISO8601 date and time.
primaryBreed	icarBreedIdentifierType		ICAR Breed code for the animal. Identifies a breed using a scheme and ID. Allows country or species-specific breeds that are a superset of the ICAR list.
coatColor	String		Colour of the animal's coat, using the conventions for that breed.
managementTag	String		The identifier used by the farmer in day to day operations.
Name	String		Name given by the farmer for this animal.
officialName	String		Official herdbook name.
production Purpose	iscProductionPurposeType	("Meat", "Milk", "Wool")	Primary production purpose for which animals are bred.
Status	iscAnimalStatusType	("Alive", "Dead", "Off-farm", "Unknown")	On-farm status of the animal (such as alive, dead, off-farm).
Parentage	array of iscAnimalParentageType		Array of parents.