DESIRA: Digitisation: Economic and Social Impacts in Rural Areas

Deliverable 5.1

27th November 2019

DESIRA receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 818194

Virtual Research Environment: specification report

CALL	H2020-RUR-2018-2
CALL	H2020-RUR-2018-2
WORK PROGRAMME	Topic SFS-29-2017 Socio-eco-economics - socio-economics in ecological approaches
Project ID	818194
Project web site	www.desira2020.eu
Document type	Deliverable n. 5.1
Title	Virtual Research Environment: specification report
Authors	Assante, M., Cirillo R., Dell'Amico A., Frosini L., Pagano P.
Date of creation	
Status of the document	Final
Version number	0.1 10/07/2019
	0.2 16/09/2019
	0.3 17/10/2019
	0.4 6/11/2019
	0.5 26/11/2019
	Final 27/11/2019
Internal reviewers	Gianluca Brunori, Elena Favilli
Keywords	Virtual Research Environment, Service, Software, Infrastructure
Coverage	

This document was produced under the terms and conditions of Grant Agreement No. 818194 for the European Commission. It does not necessarily reflect the view of the European Union and in no way anticipates the Commission's future policy in this area.

Table of contents

DESIRA receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 818194

1. Introduction

The DESIRA Virtual Research Environment specification consists of a hardware layer and a service layer. The former is made available by the D4Science Infrastructure[1] [1, 2] and it is organized as a dynamic pool of virtual machines, supporting computation and storage. The operations and management of those resources is performed via a set of ***enabling technologies*** selected to ensure availability and reliability of the infrastructure while guaranteeing reduction of costs of ownership and a set of **supporting technologies** selected to ensure secure monitoring, alerting and provisioning. The service layer, illustrated in Figure 1, consists of three service frameworks, which can be summarized as follows:

- **Enabling Framework**: the enabling framework, based on the gCube System [3], includes services required to support the operation of all services and the VREs supported by such services. As such it includes: a resource registry service, to which all e-infrastructure resources (data sources, services, computational nodes, etc.) can be dynamically (de)registered and discovered by user and other services; Authentication and Authorization services, as well as Accounting Services, capable of both granting and tracking access and usage actions from users; and a VRE manager, capable of deploying in the collaborative framework VREs inclusive of a selected number of "applications", generally intended as sets of interacting services;
- **Storage Framework**: the storage framework includes services for efficient, advanced, and on-demand management of digital data, encoded as: files in a distributed file system, collection of metadata records, and time series in spatial databases; such services are used by all other services in the architecture, exception made for the enabling framework;
- **Collaborative framework**: the collaborative framework includes all services deployed for the scientists and for each of them provides social networking services, user management services, shared workspace services. In addition, it comprises the part on the Web UI access to the Virtual Research Environment.

[1] http://www.d4science.org

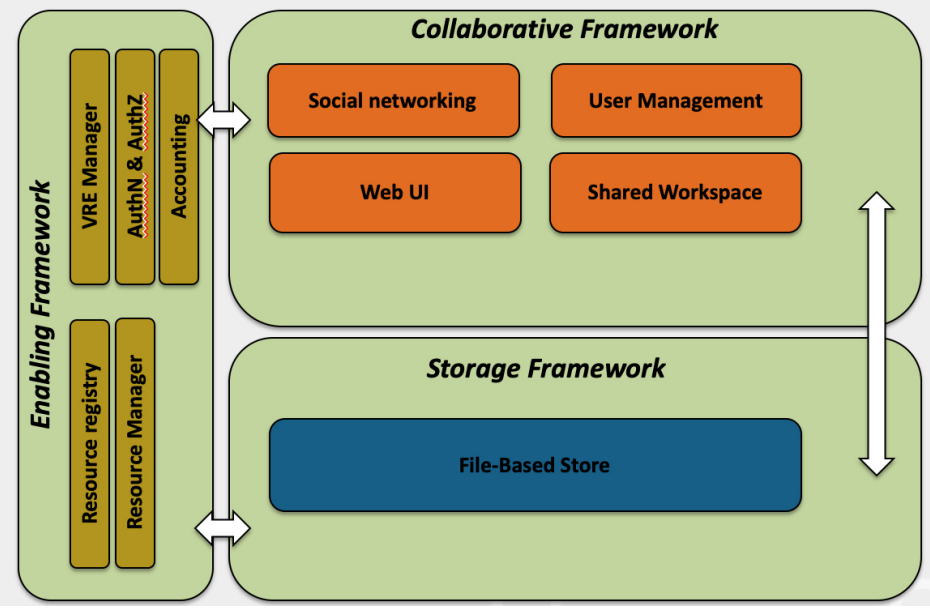

Figure 1: High-level view of the DESIRA Virtual Research Environment Specification

1.1. Structure of this report

This report is structured in two sections: Section 1 in the introduction and Section 2 is about the Virtual Research Environment Specification, Section 3 lists the references and concludes the report. More in detail, Section 2 is organized in three subsections: Enabling Framework, Storage Framework, and Collaborative Framework.

2. Virtual Research Environment Specification

The DESIRA Virtual Research Environment consists of a hardware layer and a service layer. The former is provided by the D4Science Infrastructure[2] and it is organized as a dynamic pool of virtual machines, supporting computation and storage. The operations and management of those resources is performed via a set of ***enabling technologies*** selected to ensure availability and reliability of the infrastructure while guaranteeing reduction of costs of ownership and a set of **supporting technologies** selected to ensure secure monitoring, alerting and provisioning.

The services layer is organized into layered software frameworks that increasingly hide the complexity of the cloud-based infrastructure.

2.1. Enabling Framework

2.1.1. Overview

The Enabling Framework is realized by a combination of services and libraries powered by the D4Science Infrastructure and based on the gCube System [3]. Those services promote the optimal exploitation of the resources available in the D4Science and the integration of technology external to it. They insulate as much as possible the management of the infrastructure from the data and the data management services that are hosted in or accessible through the infrastructure itself.

The motto at the heart of the management facilities is less dependencies for more management meaning that the requirements posed to resources (even independent resources) to be managed are minimal, close to zero in some cases. All the implemented solutions are prioritized in order to pursue this goal.

Towards new directions of openness and interoperability called by our growing community, management facilities move along:

- adoption of standards;
- support for new software platforms by implementing a zero-dependency approach to software management.

The Enabling framework is composed by three main systems: **Resource Management System**, **Information System**, **and Security System**. These are complex ICT systems that exploit tailored persistence technologies managed via web services.

The Resource Management System supports the creation of a Virtual Research Environment and its exploitation via the registration, management, and utilization of the resources assigned to it.

The Information System supports the registration, discovery, and access of the resources profile.

[2] http://www.d4science.org

The Security System ensures the correct exploitation, auditing, and accounting of the resources under the policies defined at registration time and customized at VRE definition time. It is orthogonal to all services operating in the infrastructure and its components are deployed on all computing nodes.

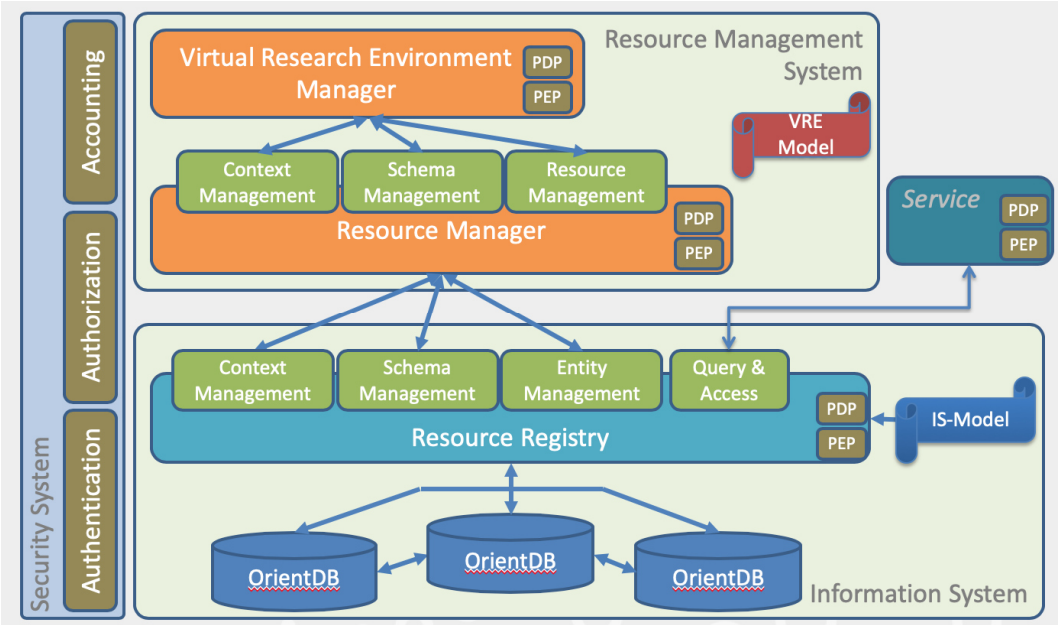

Figure 2: Enabling Framework Specification

2.1.2. Resource Management System

Within this system the **Resource Manager** is responsible for providing Resources compliant with the VRE Model. In fact, this service is the only one entitled to perform operations on the Resource Registry. It does so by exposing three port types:

1. Context Management enables Resource Registry context management by checking if the requester has the proper role/rights to do the requested action.
2. Schema Management enables schema management on Resource Registry by checking if the requester has the proper role/rights to do it;
3. Resource Management: enables to manage Resource instances by checking if:

- the requester has the proper role/rights to do the requested action;
- the action can be performed looking at the policies attached to the entities and relation instances;
- the action involves other entities or relations.

When all these checks are performed, and if and only if the action is feasible, the Resource Manager translates the incoming request in one or more outgoing requests to the Resource Registry service.

The **Virtual Research Environment (VRE) Manager** is responsible for providing context guarantees based on the VRE Model. The VRE Manager operates on the D4Science Infrastructure by using components of:

- the enabling technologies such as Resource Manager;
- supporting technologies such as Provisioning System.

The VRE Manager contacts the Resource Registry (explained in the following) to get a current view of the infrastructure; uses the provisioning system to deploy/undeploy services and data; asks the Resource Manager to update the infrastructure state consistently.

2.1.3. Information System

Within this system the **Resource Registry** is the core subsystem connecting producers and consumers of resources. It acts as a registry of the infrastructure by offering global and partial views of its resources and their current status and notification instruments. The design of the Resource Registry supports distribution and replication wherever it is possible while abstracting clients from the deployment scenario.

It exploits HAProxy for proxing requests to the deployed instances of the Resource Registry web service. HAProxy is a free, very fast and reliable solution offering high availability and load balancing for very high traffic web applications. Over the years it has become the de-facto standard open-source load balancer and it is now shipped with most mainstream Linux distributions. For these reasons, it is deployed by default in the D4Science Infrastructure.

In this system another large part is played by the database Orient DB, an open source NoSQL database management system written in Java, which contains the registry data and it is deployed as a cluster of nodes supporting different DB models such as graph, document, key/value, and object models.

2.1.4. Security System

The goal of the Security System is to protect the infrastructure resources from unauthorized accesses.

Within this system the (Service Oriented) **Authorization** and **Authentication** services are based on standard protocols and technologies, providing:

- an open and extensible architecture;
- interoperability with external infrastructures and domains, obtaining, if required, also so-called "Identity Federation";
- total isolation from the enabling framework and technologies: zero dependencies in both the directions.

The Policy-oriented Security Facilities are powered by the D4Science Authorization framework, which is a token-based authorization system. The token is a string generated on request by the Authorization service for identification purposes and associated with every entity interacting with the infrastructure (users or services).

The token is passed in every call and is automatically propagated in the lower layers.

The token can be passed to a service in 3 ways:

- using the HTTP-header: adding the value ("gcube-token","{your-token}") to the header parameters;
- using the query-string: adding gcube-token={your-token} to the existing query-string;
- logging via the default authentication widget showed by the browser using your username as username and your token as password.

The personal token can be retrieved using the token widget deployed on every environment of the DESIRA infrastructure Gateway.

This framework is compliant with the Attribute-based access control (ABAC) that defines an access control paradigm whereby access rights are granted to users through the use of policies which combine attributes together.

ABAC defines access control based on attributes that describe:

- the requesting entity (either the user or the service);
- the targeted resource (either the service or the resource);
- the desired action (read, write, delete, execute);
- the environmental or contextual information (either the VRE or the VO where the operation is executed).

ABAC is a logical access control model that is distinguishable because it controls access to objects by evaluating rules against the attributes of the entities (requesting entity or target resource) actions and the environment relevant to a request. ABAC relies upon the evaluation of attributes of the requesting entity, attributes of the targeted resource, environment conditions, and a formal relationship or access control rule defining the allowable operations for entity-resource attribute and environment condition combinations.

The Authorization framework is compliant with the XACML reference architecture. XACML is the OASIS standard for fine-grained authorization management based on the concept of Attribute-based access control (ABAC), where access control decisions are made based on attributes associated with relevant entities while operating in a given operational context, a natural evolution from Role Based Access Control (RBAC).

The **Accounting** service is also part of the security system: Accounting is defined as the recording, summarizing, and classifying of service invocations and other events, e.g. storage of data, systematically. Accountancy, in a simpler sense, is the procedure of communicating and translating raw data from the infrastructure operation to its managers and stakeholders.

This service has been explicitly developed over a storage technology and relies on the Resource Registry to discover the information needed to connect to the underling storage. In other words, any component does not have hard-coded connection information or a local configuration file. This approach allows to retrieve the storage connection information by specifying the underlying storage technology and the enabling component to use.

Figure 3 shows an example of data that can be extracted from the accounting service, specifically it shows the number of user accesses to the DESIRA Virtual Research Environment since its inception in June '19.

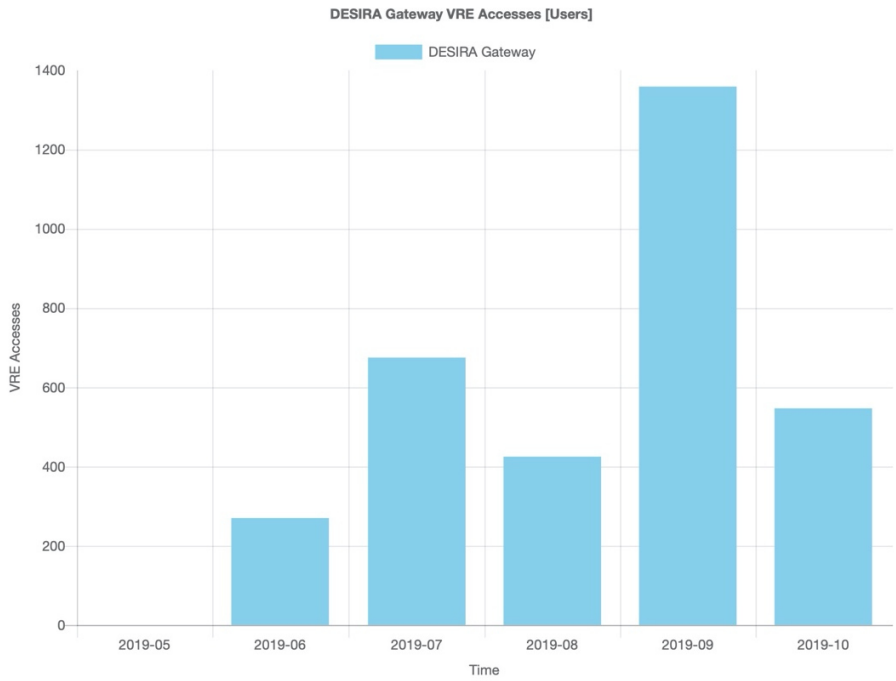

Figure 3: an example of data extracted from the accounting service

2.2. Storage Framework

2.2.1. Overview

The Storage framework is realized by a combination of services and libraries powered by the D4Science Infrastructure. It is based on the File-Based System that acts as main driver for clients interfacing the storage resources managed by the system or accessible through facilities available within the system.

The File-Based System supports functions for **standards-based** and **structured access** and **storage of files** of arbitrary size.

2.2.2. File-Based Store System

The File-Based Store system includes services providing clients functions for standards-based and structured access and storage of files of arbitrary size. This is a fundamental requirement for a wide range of system processes, including indexing, transfer, transformation, and presentation. Equally, it is a

main driver for clients that interface the resources managed by the D4Science infrastructure or accessible through facilities available within the same infrastructure.

The File-Based System is composed by a service abstracting over the physical storage and capable of mounting several different store implementations, (by default clients can make use of the MongoDB store) presenting a unified interface to the clients and allowing them to download, upload, remove, add and list files or unstructured byte-streams (binary objects). The binary objects must have owners and owners may define access rights to files, allowing private, public, or shared (group-based) access.

All the operations of this service are provided through a standards-based, POSIX-like API which supports the organization and operations normally associated with local file systems whilst offering scalable and fault-tolerant remote storage

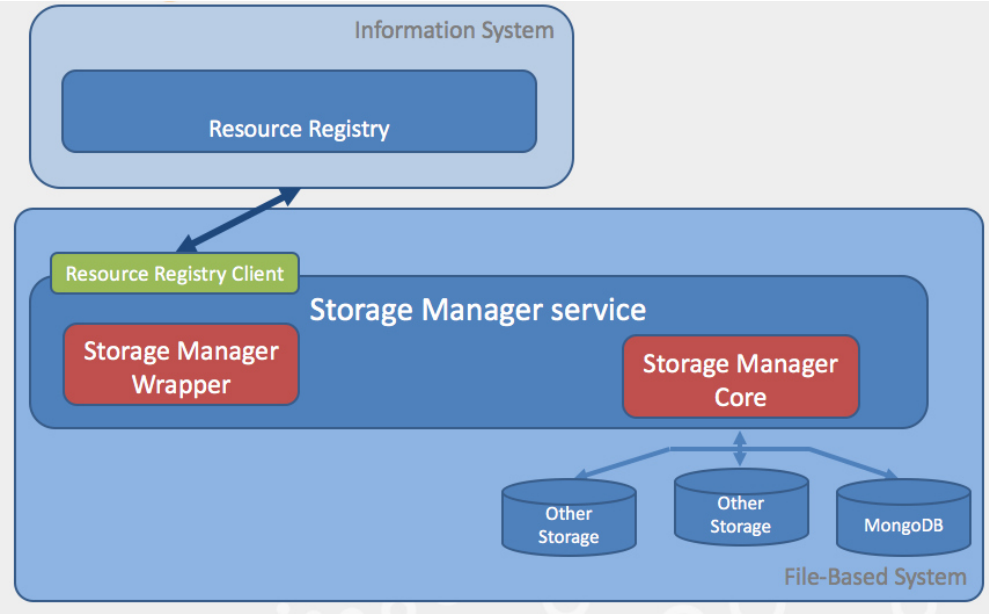

Figure 4: File-Based System Architecture

As shown in Figure 4: File-Based System Architecture the core of the Storage Manager service is a software component named Storage Manager Core that offers APIs allowing to abstracts over the physical storage. The Storage Manager Wrapper instead is a software component used to discover back-end information from the Resource Registry service of the D4Science Infrastructure. The separation between these two components is necessary to allow the usage of the service in different contexts other than the DESIRA.

2.3. Collaborative Framework

2.3.1. Overview

The Collaborative Framework is realized through a combination of software components (services and libraries) powered by the gCube System. As depicted in Figure 1, four main subsystems characterise the Collaborative Framework:

- Social Networking;
- Shared workspace;
- User Management;
- Web UI.

The first three systems provide consumers with a homogenous abstraction layer over different external technologies enabling to operation of the framework. The external technologies involved comprises, Apache Cassandra, Apache Jackrabbit, Elastic Search, MongoDB, and Liferay Portal. In particular, the Social Networking System exploits an Apache Cassandra cluster and an Elastic Search cluster, the Shared Workspace System exploits an Apache Jackrabbit repository (metadata) and a MongoDB cluster (payload) for its backend, the User Management System exploits Liferay Portal for its backend and to allow users to login for personalized services or views. The latter system, the Web UI, is composed by a set of software libraries to realise a Liferay based web-portal customized to interface with the D4Science infrastructure and equipped with Java portlets. Through it, users have access to the resources and Virtual Research Environments created to serve the needs of the DESIRA community and scenarios.

2.3.2. Social Networking

Social Networking System comprising services conceptually close to the common ones promoted by social networks – e.g., posting news, commenting on posted news, likes, private messages and notifications; It is composed by 2 main services, the Social Networking Service and the Social Indexer Service.

The Social Networking Service logic relies on the Social Networking Model, this Model is used also for the efficient storage of the Social Networking Data (Posts, Comments, Notifications etc.) in the underlying Apache Cassandra Cluster. This Cluster is queried by means of a Java client.

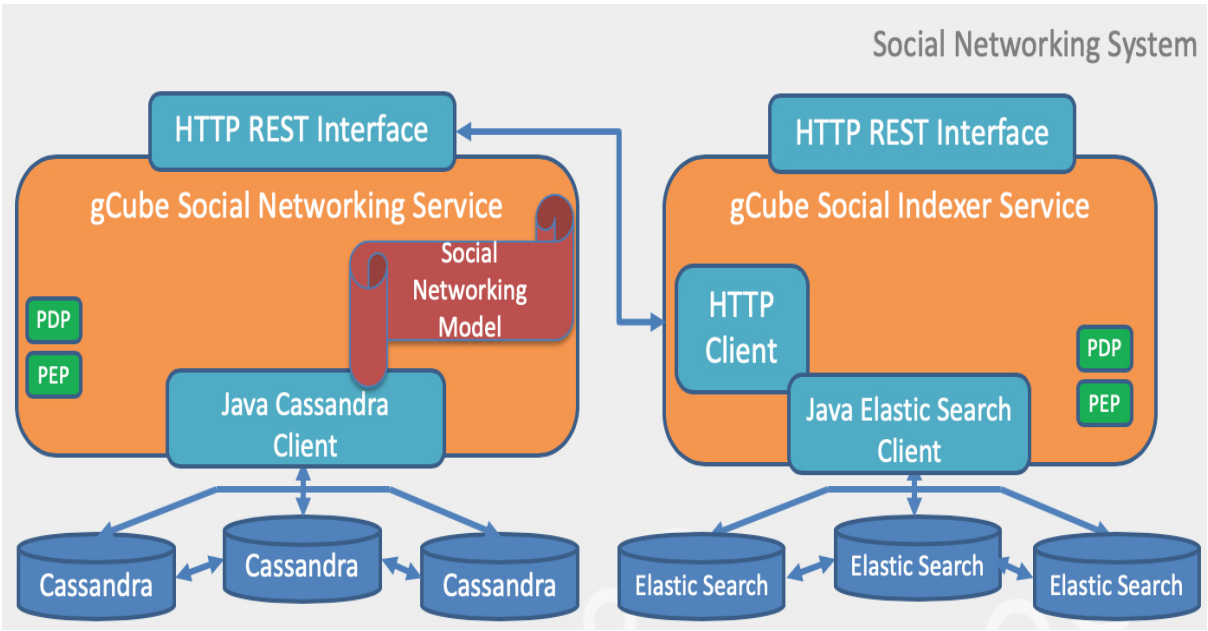

Figure 5: Social Networking System Architecture

The Social Networking Service exposes an HTTP REST Interface for the external, and non, services of the infrastructure. The Social Indexer Service uses such interface for the retrieval of the Social Networking Data to index by means of an Elastic Search Cluster. The Social Indexer Service exposes an HTTP REST Interface for the external, and non, services of the infrastructure needing to perform search operations over the Social Networking Data.

Both Services rely on the Policy Decision Point (PDP) and the Policy Enforcement Point (PEP) to intercepts user's access request and evaluate these requests against authorization policies of the Authorization System of the Infrastructure.

2.3.3. Shared Workspace

The Shared Workspace System provides a remote (Cloud) folder-based file system, supporting sharing of folders and different item types (ranging from binary files to information objects representing, for instance, tabular data, workflows, distribution maps, statistical algorithms).

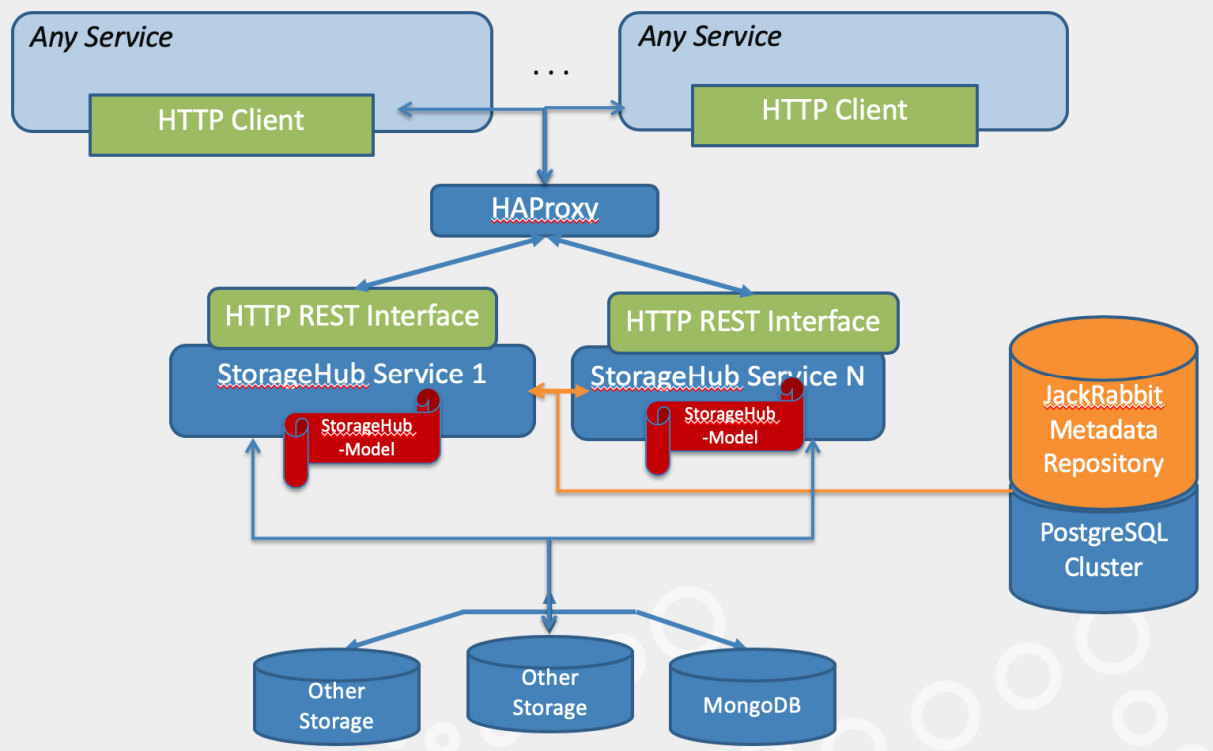

Figure 6: Shared Workspace System Architecture

The Shared Workspace System consists of one D4Science service, named StorageHub Service, relying on 2 different storage technologies to store the metadata of the items being stored, namely Apache Jackrabbit as metadata repository and PostgreSQL as Apache Jackrabbit Bach-end Database.

The StorageHub Service is replicable and a HAProxy on top is used for proxing requests to the deployed instances of it. One other distinguish feature of the StorageHub Service is that the actual payload of the items can be stored on a number of in house and commercial storage technologies, for instance in a MongoDB Cluster, but also on other types, including Cloud Storages solutions (e.g. Amazon S3).

The StorageHub Service identifies a core set of capabilities to work on JackRabbit content. Together with its model, named StorageHub model, exposes content in the content repository as HTTP resources, fostering a RESTful style of application architecture. Home RESTFUL interface processes HTTP requests coming from clients. The following operations are supported:

- retrieve content;
- create content;
- modify existing content;
- remove existing content;
- move existing content to a new location;
- copy existing content to a new location.

2.3.4. User Management System

Users are the fundamental entity managed by this System. As a matter of fact, User is an entity that can sign into the portal and do something. Users are assigned a Role and a Role is what defines the user's privileges. The User Management System provides functionality to manage personal profiles and users in the DESIRA Virtual Research Environments, supporting user groups (for the purpose of group specific privileges) and roles for application specific needs related to the user's role in DESIRA Project.

2.3.5. Web User Interface Access

The Web UI is composed by a set of software libraries to realise a Liferay based web-portal customized to interface with the D4Science infrastructure and equipped with Java portlets. Through it, users have access to the resources and Virtual Research Environments created to serve the needs of the DESIRA community and scenarios.

DESIRA Infrastructure Gateway

The DESIRA Infrastructure gateway is expected to be the end-user access point to the DESIRA services and Virtual Research Environments. A screenshot showing the DESIRA VREs offering, at the time of writing, is given in Figure 7.

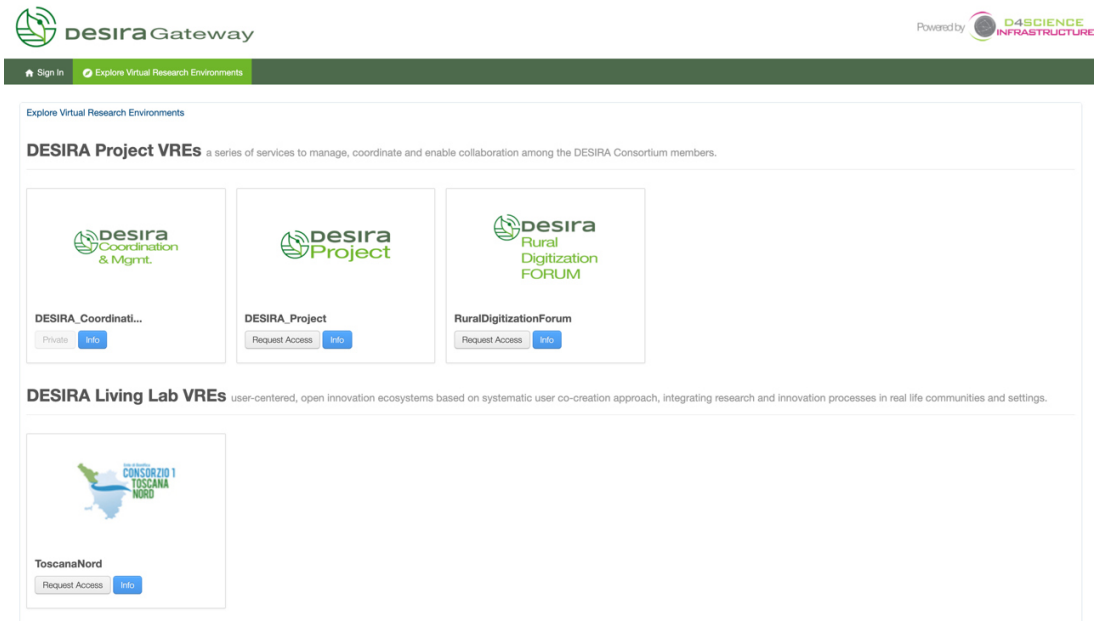

Figure 7: DESIRA VREs offering on the Infrastructure gateway VREs

Users are provided with log-in facilities and once logged in they are provided with a dashboard (Figure 8) realising a user-friendly working environment. As shown in Figure 8, the dashboard provides users with a personalised view of the whole DESIRA offering in one single place.

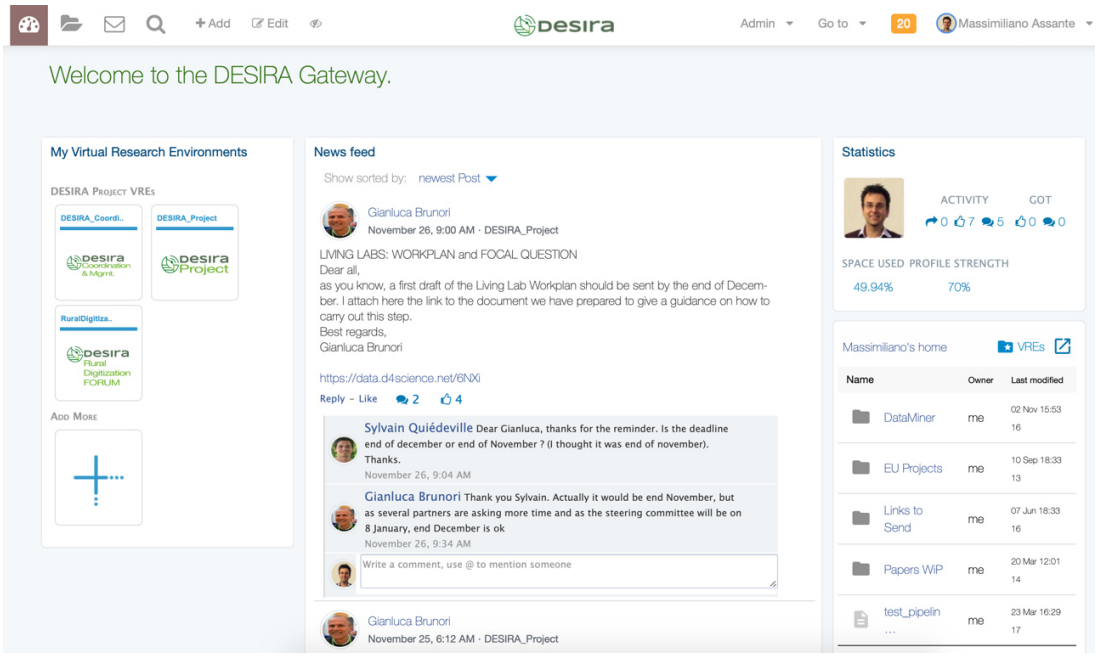

Figure 8: DESIRA Infrastructure gateway Dashboard

The Dashboard page contains (from the left) *(i)* the list of VREs a user belongs to, *(ii)* a News feed gathering all the posts created in the VREs a user belongs to sorted in reverse chronological order (newest first), *(iii)* some statistics about the activity performed by the user and *(iv)* a small view on the user's shared workspace (cf. ref 2.3.3) file and folders.

Moreover, the Dashboard top bar contains quick links to access the following facilities:

- User Shared Workspace;
- Private Messages;
- Search posts;
- User Profile and Account.

DESIRA Virtual Research Environments Web UI

At the time of writing this document four (4) Virtual Research Environments (VREs) have been created and are operational. These VREs have been classified following the offering type, namely DESIRA Project and DESIRA Living Labs.

• **DESIRA Project VREs**: these VREs are dedicated to the management, coordination and enabling of collaboration among the DESIRA Consortium member. At the time of writing 3 VREs are available for this class, namely:

1. DESIRA Project: conceived to be the working environment supporting cooperation and collaboration among the DESIRA project beneficiaries;
2. DESIRA Coordinantion & Management: used by the management and coordinators of DESIRA Project;
3. Rural Digitization Forum: a EU-level network composed of representatives from the Living Labs, Rural businesses and services, Public Authorities, Citizen groups, Digital technology operators, Farmers, Media, Scientists, and representatives of other relevant projects.

• **DESIRA Living Labs VREs**: these VRE will be user-centered, open innovation ecosystems based on systematic user co-creation approach, integrating research and innovation processes in real life communities and settings. At the time of writing only one VRE of this type is available, namely "Toscana Nord".

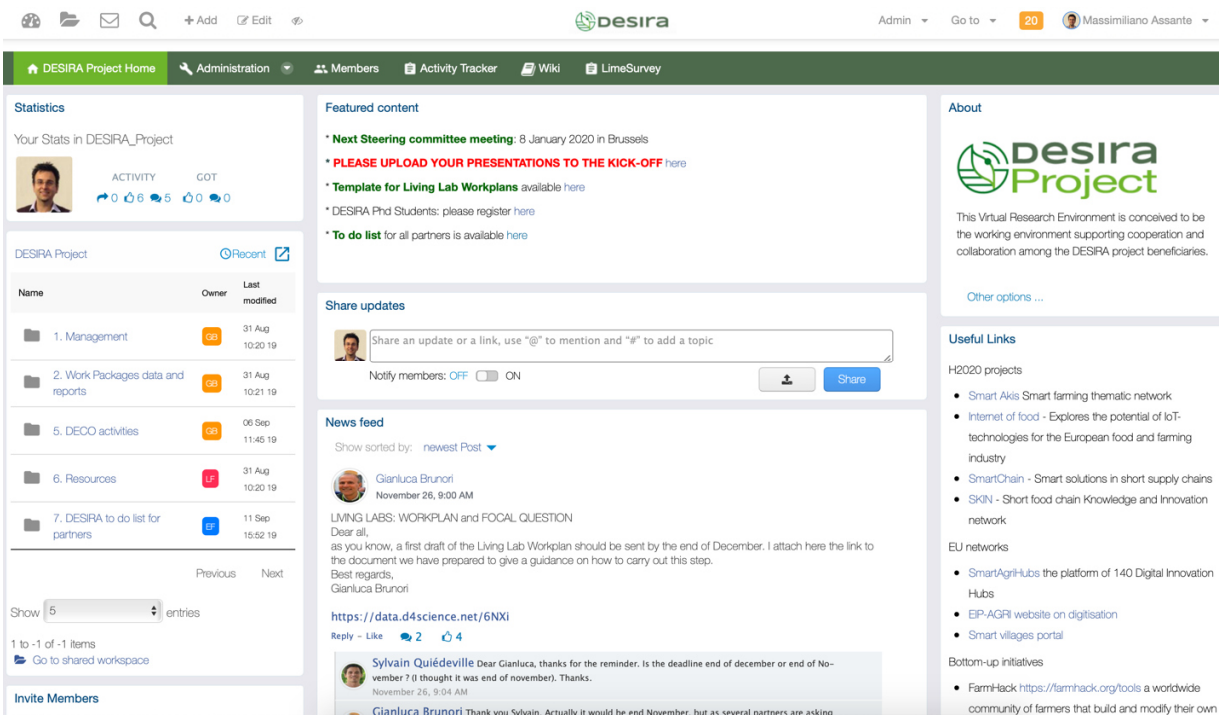

Figure 9: DESIRA Project VRE Home

Figure 9 shows a screenshot of the DESIRA Project VRE Home, conceived to provide DESIRA project members with a VRE-based working environment equipped with tools (conversations, wiki, activity tracker etc.) useful to the project management.

The main functionality offered by the VRE include:

- A dashboard (cf. Figure 9) containing a Featured Content part and applications for (i) collaborating with other VRE members by posting messages or being informed and reacting to (e.g. commenting, favouring) co-workers' posts, (ii) being informed on the most recent objects added to the VRE workspace, (iii) invite colleagues to participate, (iv) have information on the VRE including a description and the list of managers, and (v) having some indicators on the user activity in the VRE and useful Links;
- A Wiki for collaboratively documenting project related activities;
- An Activity Tracking system (or ticketing system), to support the planning and monitoring of project related activities including work packages, tasks, deliverables and milestones as well as technology development tasks;
- A shared area in the workspace, for making available objects of interests, e.g. project deliverables, presentations, working notes;
- A Members area, for enabling each VRE member to be informed on the rest of VRE members and acquire details for contacting them;
- A user management area, to enable authorised users (i.e. VRE Managers) to manage other users using or willing to access the VRE. VRE Managers can (i) authorise users in accessing the VRE and its services, (ii) assign or withdraw roles to users, (iii) remove users, and (iv) send a communication to the current users;
- An accounting area, to enable authorised users (e.g. VRE Managers) to analyse usage records pertaining VRE services, e.g. most used service, most active user.

REFERENCES

[1] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, F. Sinibaldi, **Enacting open science by D4Science**, Future Generation Computer System (2019) DOI: 10.1016/j.future.2019.05.063

[2] L. Candela, D. Castelli, P. Pagano (2013) **Virtual Research Environments: An Overview and a Research Agenda**. Data Science Journal, Vol. 12

[3] Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., Panichi, G., Perciante, C., Sinibaldi, F. **The gCube System: Delivering Virtual Research Environments as-a- Service**. Future Generation Computer Systems (Vol. 95) DOI: 10.1016/j.future.2018.10.035