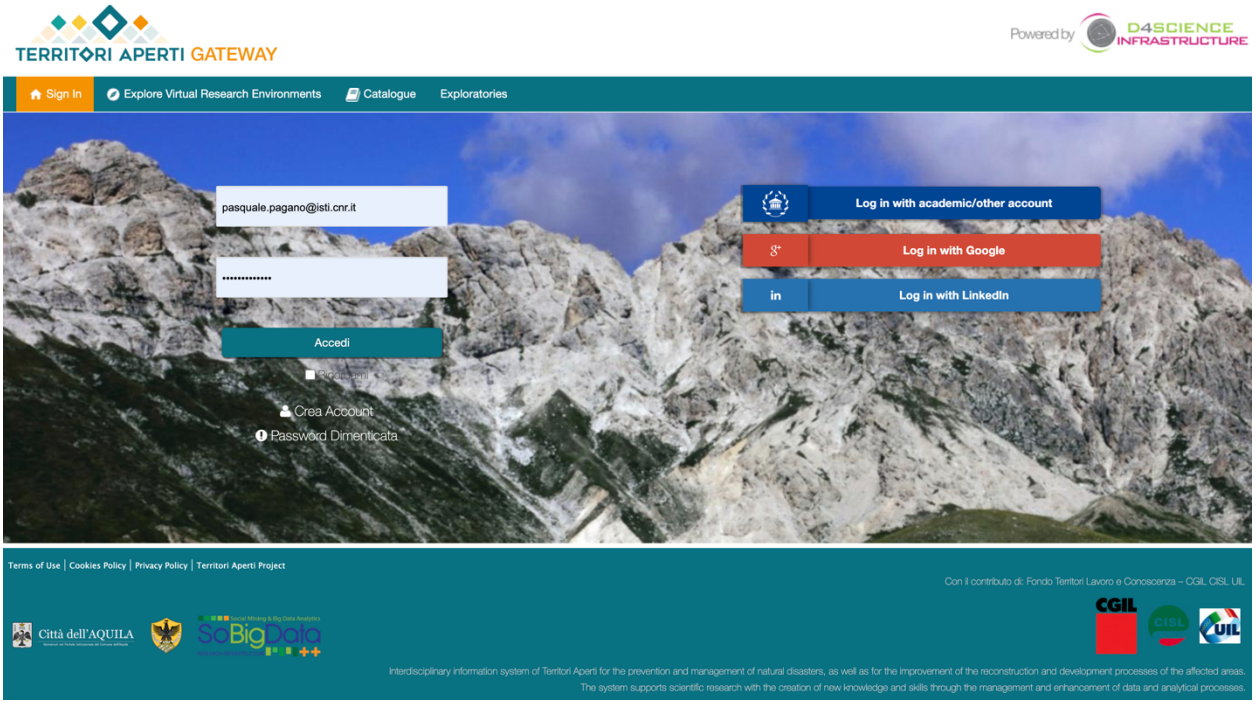
Territori Aperti

Report N° 1



30.11.2020

Pasquale Pagano, Massimiliano Assante, Andrea Dell'Amico

D4Science

V 1.0

Date	Description	Author	Version
30.11.2020	Final Report	CNR	1.0

1 Introduction

The Territori Aperti Gateway, https://territoriaperti.d4science.org/, provides access to the Territori Aperti Catalogue and to the Exploratory that support scientific research with the creation of new knowledge and skills through the management and enhancement of data and analytical processes.

This document illustrates the services exploited by the Territori Aperti Gateway and reports the status of the activities at the end of the first 4 months ed in particular the exploitation of the VREs.

2 Services

The Territori Aperti Gateway is equipped with the following generic facilities:

- A **gateway service** to provide users with a web portal to access the VREs.
- A shared **workspace service** to enable every user to store and organise the information objects he/she is interested in working with. In addition to that, the user is allowed to collaborate with other users by sharing objects and messages. To every user is guaranteed 100 GB storage volume;
- A **publishing platform** to support data harmonization and publication. It resembles a catalogue of artefacts with search and browse, yet the openness with respect to the typologies of products published, the metadata to document them as well as the integration with the rest make it a flexible environment
- A **social networking collaboration platform** to enable users to use the common facilities typical of social networks – e.g., posting news, commenting on posted news – yet adapted to the settings of the working environments. Users can post news as well as applications. This platform embeds the following services:
    - A **messaging service** to provide users with a common email environment as-a-Service. The distinguishing feature is represented by its integration with the other services, e.g., it is possible to send any information object residing in the workspace (regardless of how “big” and “complex” it may be) as an attachment without consuming bandwidth;
    - A **notification service** to alert users on relevant activities as they happen. These notifications offer a sense of anticipation and create a productivity boost. Users receive an alert (through a priori selected channels, e.g., email, web portal, twitter) notifying them when something of interest has happened in their VRE(s); The notification service is equipped with a **member facility** to provide users with a list of the VRE co-workers, i.e., the list of members partaking in the VRE and contributing to it;

- A **VRE Management service** to enable authorized users (i.e. VRE Managers) to manage other users using or wanting to access the VRE. VRE Managers can (i) authorize users for access to the VRE, (ii) assign or withdraw roles to users, (iii) remove users, and (iv) send communications to the current users.

The Gateway exploits resources maintained and operated of the D4Science Infrastructure.

Upon registration to the **Gateway Service**, the user can immediately use the storage space via the **Workspace Service**, the **Social Networking Collaborative Platform** including the ***Email Service*** to send/receive data to other registered users, the ***Social Service*** to share and read news and posts with your connections and the ***Notification Service*** for user notifications, and the **Catalogue Service** to browse available datasets, methods, and services therein published. The can also apply to one or more moderated or public **Virtual Research Environments** offering one or more additional services.

All the Services made accessible through this Gateway are also accessible through APIs by specifying the secure token generated with the registration and specialised for each Virtual Research Environment the user is member of. Starting from January 1st 2020 the APIs are exploitable only by using the secure and encrypted communication protocol over a computer network, HTPPS, while the support for the deprecated communication protocol over a computer network, HTTP, is definitely terminated December 31st 2019.

2.1 Gateway Service

The Gateway Service enables registered users to access Gateway services and one or more VREs. The Gateway Service employs the Liferay portal as the portlet-hosting platform. Liferay Portal offers a complete platform for building web apps, mobile apps, and web services quickly, using features and frameworks designed for rapid development, good performance, and ease of use. It runs on all major application servers and servlets containers and it is JSR 168 and JSR 286 compliant. Portlets are using JSR 286 and several other technologies such as Java Server Pages for dynamically generation of HTML/XML documents in response to a client's request, and the most popular Front-End frameworks such as Angular, React or Vue.

2.2 Workspace Service

The Shared Workspace is an online environment to support secure and controlled data storage and sharing.

This facility relies on Apache Jackrabbit for storing and managing workspace item representations. Items payload are stored by relying on a hybrid cloud storage solution that, by means of ad-hoc plugins, exploits various storage solutions suitable for diverse typologies of content, e.g. MongoDB for binary files.

Every workspace item is

- equipped with an actionable unique identifier that can be used for citation and access purposes;
- versioned and a new version is automatically produced whenever the item is explicitly changed by a user or any application/service of the VRE on behalf of an authorised user;
- equipped with rich and extensible business metadata that capture descriptive features as well as lineage features
- organised in folders that can be
    - private: content is available only for its owner;
    - shared: content is available for selected users (decided by the owner);
    - VRE folder, content is available to VRE members;

The Workspace is tightly integrated with both the social networking and the catalogue for easing the dissemination of its artefacts.

Data Managers can exploit the Shared Workspace while keeping accounting and traceability. Individual scientists can easily share data, experiments, derived data products with selected colleagues. Working groups empower the working group with an easy-to-use environment where operations are traced and accounted.

A backup and recovery strategy protect about data loss incidents that can take a variety of forms, but when it comes to a workspace technology, they generally fall into two categories: catastrophic failure and human error. Even if catastrophic failure includes natural disasters and any other scenario that permanently destroy all the server nodes powering the Workspace service, most of the issues are generated by human error. Humans introduce application bugs or accidentally delete data. A bad code release that corrupts some or all of the production data is an unfortunate but common example. In the case of human error, the errors introduced will propagate automatically to the replicas, often within seconds.

The Workspace service provides continuous, online backup for data as a fully managed service. The service streams encrypted and compressed *oplog* data to a dedicated server so that a continuous backup is activated. By default, the Workspace takes snapshots every 6 hours and oplog data is retained for at least 24 hours and in any case till all the servers of the storage cluster are properly synchronized.

2.3 Catalogue Service

The Publication Platform is a comprehensive environment to support data harmonization and

publication. It resembles a catalogue of artefacts with search and browse, yet the openness with respect to the typologies of products published, the metadata to document them as well as the integration with the rest make it a flexible environment. Every published item in the catalogue is characterised by *(i)* a type, which highlights its features and allows an easier search, *(ii)* an open-ended set of metadata which carefully describe the item, and *(iii)* optional resource(s) representing the actual payload of the item.

This platform primarily relies on CKAN technology, i.e. an open source software enabling users to build and operate open data catalogues. This core technology has been wrapped and extended by means of the Catalogue Service, a component realizing the business logic of the publishing platform. The Catalogue Service enacts the management of Catalogue Items. A Catalogue Portlet, accessible in each VRE, allows navigating the catalogue content as well as publishing content by exploiting the Publishing Widget. This widget is also embedded into the Workspace portlet, so users can publish folders and/or files directly from there. External services can access the catalogue content and publish new items via the gCube Catalogue RESTful APIs. The Catalogue Service relies on the Workspace for storing the payload of the published items.

Each catalogue item type carefully defines the metadata elements characterising the item typology by specifying the names of the attributes, the possible values, whether an attribute is a single instance or a repeatable one. In addition to that, each item type contains directives on how to exploit attributes for items organisational purposes, e.g. automatically transform values in tags or exploit the values for creating collections or groups of items. Moreover, the following properties apply:

- every catalogue item is equipped with an actionable, persistent, unique identifier (namely a PURL) that can be used for citation and access purposes;
- whenever a catalogue item is published, the associated payload(s) is stored in a persistent and unchangeable storage area to guarantee its long-term availability;
- every catalogue item is equipped with a license carefully characterising the possible (re-)uses;
- every publication of an item may lead (according to the VRE configuration selected by the VRE Manager) to the automatic production of a post in the social networking area of the VRE to inform its members;
- every catalogue item is equipped with rich and open metadata, i.e. it is possible to carefully customise the typologies of products and the accompanying metadata to the community needs;
- catalogue contents (item's metadata and resources) are made available for consumption by clients by the RESTful API as well as by other standard APIs, e.g. DCAT and OAI-PMH.

Data Managers extend current practices with controlled and standard data publication with the aim to enlarge the access to data products while maintained full ownership and control over the sharing of results. Individual scientists collaborate via standard formats and protocols with selected colleagues since the conception, definition, validation, and sharing of a result. Working groups normalize the exchange of data on activities, enlarge the audience of products, user and data management for the working group.

The D4Science Infrastructure offers data sharing at different levels: within a Catalogue accessible only in the VRE, accessible in different VREs and publicly accessible via the Gateway service.

2.4 Social Networking Collaborative Platform

The Social Networking Collaborative Platform enables a comprehensive and collaborative environment to support sharing and collaboration among users. It resembles a social networking environment with posts, tags, mentions, comments and reactions, yet its integration with the other services makes it a powerful and flexible communication channel for researchers.

2.4.1 Social Networking Service

The communication about Virtual Research Environment users relies on the Social Networking Engine, a Cassandra database for storing social networking related data and on Elasticsearch for the retrieval of social networking data. The Engine exposes its facilities by an HTTPS REST Interface and comprises two services: (i) the Social Networking Service that efficiently stores and accesses social networking data (Posts, Comments, Notifications, etc.) in the underlying Cassandra Cluster, and (ii) the Social Networking Indexer Service that builds Elasticsearch indices to perform search operations over the social networking data.

There is no predefined way to structure a discussion; users can start new discussion threads, annotate them with tags for easing the cataloguing and discovery, refer to other threads and material both internally stored and available on the web. Every (re)action performed by a user – be it a new post, a reply to a post, or the rating of a certain post or post reply – is carefully captured, documented, and equipped with an actionable unique identifier that can be used for citation and access purposes.

Individual scientists can easily share ideas, comments, suggestions and communicate with other members of the community. Trainers can easily share data, algorithms, and technologies and communicate with the classroom. Trainees can easily access to data and technologies in a controlled environment where experiments and tests can be performed, traced and documented.

2.4.2 Messaging Service

The Social Networking Collaborative Platform implements a Messaging Service, in order to enable messages exchange only between the D4Science Infrastructure registered users.

This facility relies on the Social Networking service and on Apache Cassandra for storing, indexing and managing messages.

2.4.3 Notification Service

For the purpose of facilitating the user interaction, the Social Networking Collaborative Platform includes a Notification Service. The Notification Service can be configured from the Gateway to an email address associated with the user account.

2.5 Virtual Research Environments (VREs)

In the D4Science Infrastructure, services and related access to data are available through Virtual Research Environments (VREs). These VREs can take various forms including web interactive user interface, web applications, pluggable standalone user interface.

VREs can be created for a specific period and a specific task and only authorized users access data and services exposed through these VREs.

There are different levels of user authorization available: VRE Manager, VRE Designer, VRE Data Manager and VRE User:

- VRE Designers select the services/data that are made available in the VRE;
- VRE Managers have the right to grant access to other users and to assign roles;
- VRE Data Managers have the right to register/unregister datasets and make them available to all Users of the VRE;
- VRE Users can upload and manage data within dedicated VRE(s) according to their authorization level.

Users can apply to a **Moderated** or a **Public** Virtual Research Environment:

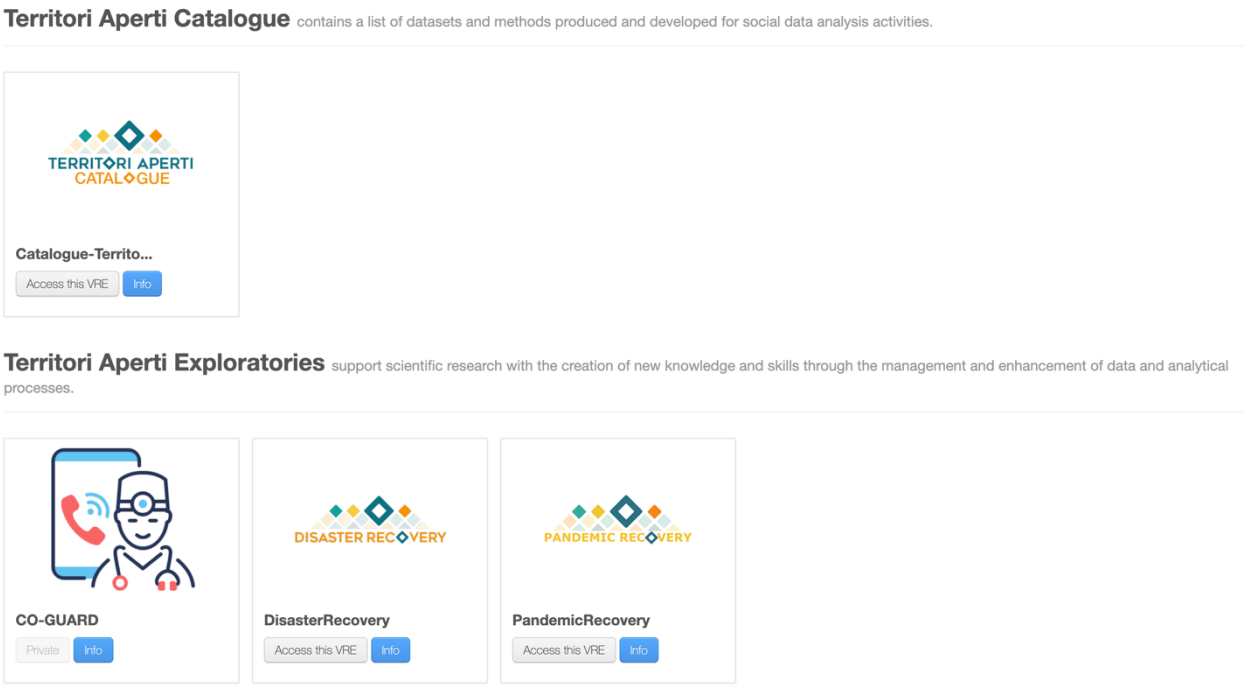
- Access to the Moderated VREs is restricted. The acceptance of an application issued by a user is solely under the responsibility of the VRE Manager that has to verify that the user belongs to the community operating the VRE.
- Access to the Public VRE is regulated by an automatic procedure that grants the access to any user applied for it.

Each VRE may offer access to a **Data Analytics Platform** to perform data analysis; a **Tabular Data Manager Platform** to collate, harmonize and manage tabular data; a **NLPHub** to orchestrate and combine several state-of-the-art text mining services that recognize spatiotemporal events, keywords,

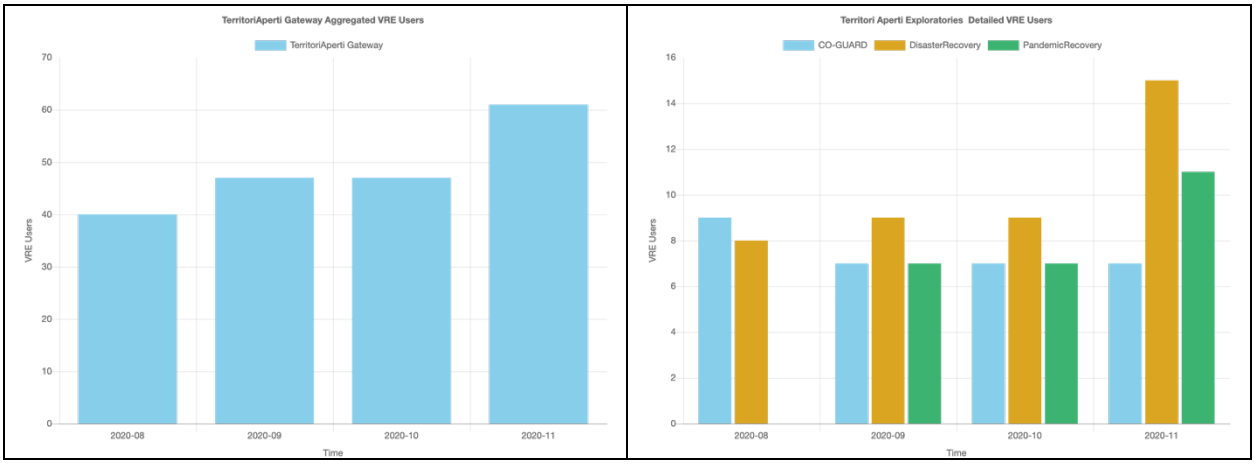
and a large set of named entities; and to third-party software as **Share-Latex** and **Shiny Proxy**.

3 Exploitation Status

The Territori Aperti gateway provides access to 4 different environments: 1 Catalogue and 3 Exploratory.



The following indicators documents the exploitation of those VREs in the first 4 months from August 2020 to November 2020. The indicators are extracted from the Accounting Dashboard [1].



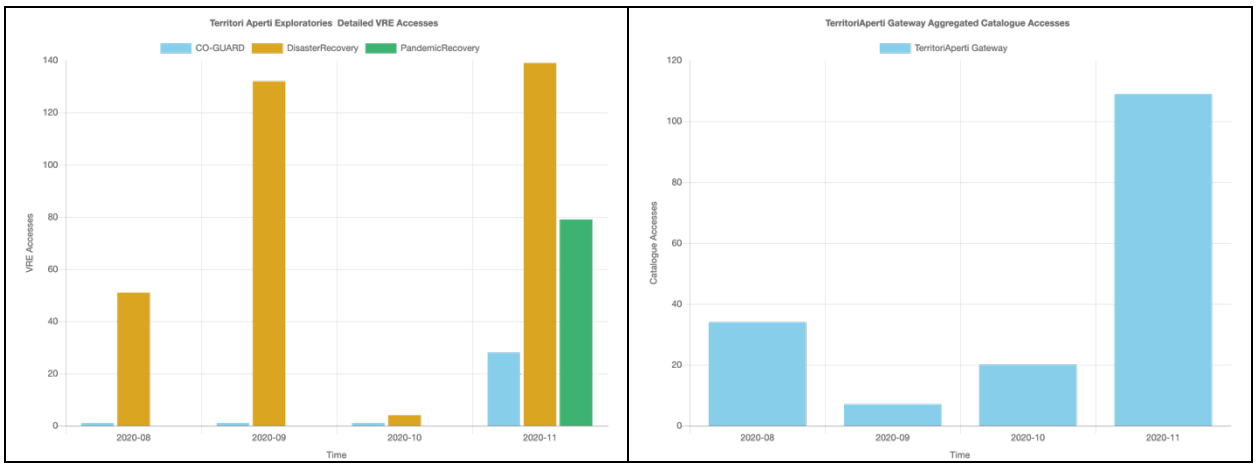


Table 1: Users and Accesses from August to November 2020

3.1 Training Report

The training event was held November 27th and it covered the first two planned topics:

1. L'infrastruttura D4science per Territori Aperti: panoramica delle funzionalità offerte agli utilizzatori. 60' (45 + 15), November 27th, 2020 14:30 - 15:30 CET - https://data.d4science.net/vZao
2. L'infrastruttura D4science per Territori Aperti: panoramica delle funzionalità di management del gateway, degli esploratori e del catalogo. 45' (30 + 15), November 27th, 2020 15:45 - 16:30 CET - https://data.d4science.net/y7G3

4 References

1. Accounting Dashboard (restricted access) https://territoriaperti.d4science.org/group/d4research/dashboard
2. Accounting raw data inspector (restricted access) https://territoriaperti.d4science.org/group/d4research/manager