Project Acronym:	**HUTER**
Project Full Name:	**Human Uterus Cell Atlas**
Call identifier:	H2020-SC1-2019-Single-Stage-RTD
Topic:	SC1-BHC-31-2019
Grant Agreement No:	874867
Start date of Project:	01/01/2020
Project Duration	2 years (24 months)
Document due date:	30/09/2020
Submission Date	30/09/2020
Leader of this report:	BAHIA
Deliverable no:	2.2
Deliverable name:	Deployment of HUTER cloud infrastructure
Dissemination level:	**Public**

Version History

Version	Date	Details
1.0	30/09/2020	Deliverable completed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874867.

DELIVERABLE 2.2

Deployment of HUTER cloud infrastructure

Table of Contents

List of Acronyms

AWS	Amazon Web Services		**BAHIA**	Bahía Software SLU		**DICOM**	Digital Imaging and Communication On Medicine
EC	European Commission		**eCRF**	electronic Case Report Form		**HCA**	Human Cell Atlas
HCA-DCP	Human Cell Atlas - Data Coordination Platform		**HUTER**	Human Uterus Cell Atlas		**IAM**	Identity and Access Management
IT	Information technologies		**PACS**	Picture archiving and communication system		**S3**	Simple Storage Service
SME	Small and medium-sized enterprise		**WDL**	Worflow description language		**WP**	Work package

1. PURPOSE OF THIS DOCUMENT

The development of the activities regarding Deliverable 2.2 which is defined as “Deployment of HUTER cloud infrastructure” have been successfully completed. The purpose of this document is to provide an overview about the work performed in the framework of this Deliverable as part of the WP2 (entitled “HUTER platform infrastructure”) that BAHIA lead. However, the complete technical definition of the HUTER Cloud infrastructure has not been included in this document because is part of the Deliverable 2.1 “Platform architecture design”.

This deliverable has been fully developed by BAHIA partner as expert SME in IT solutions considering all the HUTER project requirements.

1.1. Related documents

Documents linked to current actions to be delivered:

HUTER_WP2_D2.2_Deployment_of_HUTER_cloud_infrastructure

Documents linked to future actions to be delivered:

HUTER_WP2_D2.3_Beta_version_of_data_access_tools

HUTER_WP2_D2.4_Final_implementation_of_data_access_tools_and_DICOM_visualization_tool

HUTER_WP7_D7.1_Final_design_ of_HUTER_platform_architecture

HUTER_WP7_D7.2_Visual_System_implemented_&_Digitalisation_software_to_transform_images_from_research_equipment_under_opensource_standards

Other documents referenced:

HUTER_WP1_D1.1_Ethics_Plan

HUTER_WP9_D9.2_Data_Management_Plan

2. STATUS OF DELIVERABLE

The deployment of HUTER cloud infrastructure has been developed, so we considered the deliverable to be completed.

3. INTRODUCTION

The HUTER project is focused on creating the molecular reference map of the human uterus. For this purpose, HUTER researchers will generate vast amounts of molecular and imaging data from cell sequencing

technologies and uterus cells of women samples. In this line, HUTER will be part of Human Cell Atlas global initiative which aims to create reference cellular maps of whole-body tissues and organs. As a part of HCA global initiative, HUTER must share its generated data openly with the HCA research community with the aim of gathering all the completed cellular maps of the human body in a single platform. Although an open cloud-based Data Coordination Platform (HCA-DCP) is being built to check, share and analyse data to be generated under the different HCA projects, it is not totally finished to allow HUTER partners develop their commitments with the EC. Therefore, BAHIA is leading the development of a cloud-based hosting for HUTER project to guarantee our competence to meet our project deliverables in due time and overcome other unexpected obstacles as a contingency measure. The HUTER cloud-based hosting will be active during all HUTER project and will be turned off once the project ends but the data hosted will be gradually transferred to HCA-DCP when its construction is completed and the data sharing is authorized. Considering the provisional nature of the HUTER cloud because the persistence of data will be done in HCA-DCP servers once the HUTER project ends, the deployment of the HUTER cloud infrastructure has been based on the alignment with HCA guidelines and bearing in mind three key technical aspects: scalability, high availability and storage.

Furthermore, the HUTER cloud infrastructure will not only host complex data from cell sequencing technologies but also support data from other relevant specific components such as website for communication and dissemination activities, intranet for project managing, advanced DICOM viewer, etc. These are specific components that are not related to HCA-DCP but they are also needed to fulfil the specific HUTER project objectives and requirements. Thus, all these features and functionalities must be supported by the HUTER cloud infrastructure to a proper development of HUTER project. For this reason, the HUTER cloud infrastructure was customly designed and deployed, integrating and supporting all these required components.

In order to demonstrate that the deployment of the HUTER cloud infrastructure was successfully carried out, we will provide some evidences of different components deployed over the HUTER cloud infrastructure on this document. Some of these evidences will include not only functional descriptions of the components currently deployed but also screenshots related to these components working properly over the HUTER cloud infrastructure. We want to highlight that the first version of the HUTER cloud infrastructure design is fully defined in Deliverable 2.1 (HUTER_WP2_D2.1_Platform_architecture_design) and consequently it will not be addressed on this document.

On the other hand, some platform components are currently being designed and developed in close collaboration with partners thus it is not possible to deploy them yet. Nonetheless, the hardware infrastructure required for their deployment is ready to support them once the development is finished.

4. CURRENTLY DEPLOYED COMPONENTS

Herein, components of HUTER platform that have been successfully deployed will be briefly described below, focusing on their functional relevance. Furthermore, we will support our work including several evidences in screenshot format to justify that the committed tasks of this deliverable were properly performed. However, we want to highlight that we must assure confidentiality of data gathered in the HUTER platform complying ethical and data management agreements following European and national legislation in this regard. Therefore, it has not been included any credentials or access instructions to HUTER cloud infrastructure on this document to avoid unauthorized accessions to sensible data due to open/public nature of this document. We considered that we provided enough documented evidences to show that we have accomplished properly with our commitments regarding Deliverable 2.2. Otherwise, EC officers are welcome to contact BAHIA (proyecto.huter@bahiasoftware.es) to request further specific details.

4.1. AWS Nextwork infrastructure

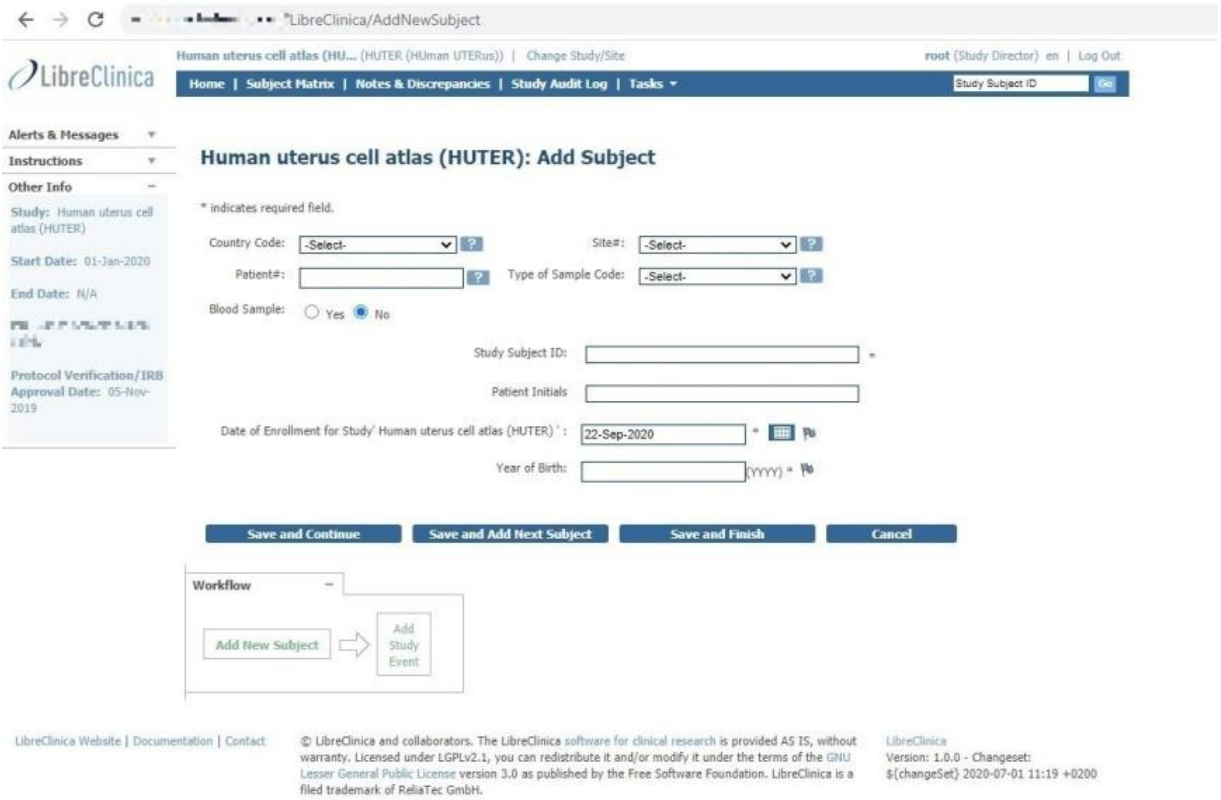
In order to get support for all next listed components, the documented AWS architecture has been deployed in Ireland data center to provide two deployment environments called “production” and “development”. The first one is addressed to support the researcher’s work, store real data and execute high performance analysis. Meanwhile, development is the environment for the deployment of new tools in order to get them tested before their exploitation.

4.2. LibreClinica

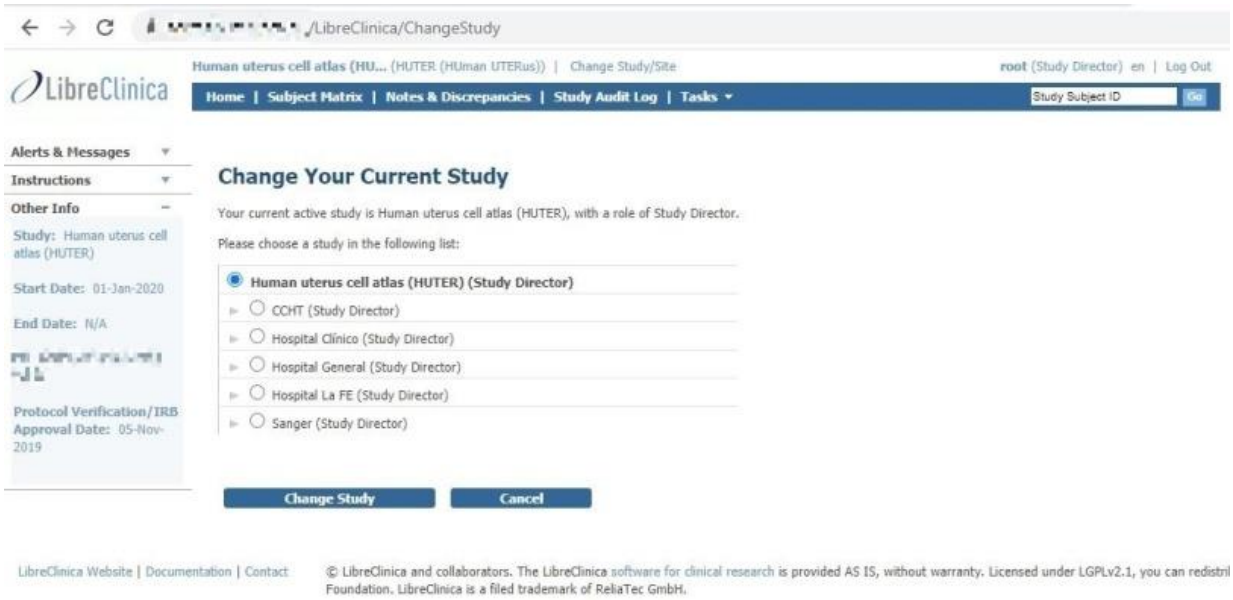
This component is an open source software initially designed for clinical trial studies but now adapted to HUTER project requirements by BAHIA. This component allows to manage electronic case report forms (eCRFs) through a user-friendly web interface that will contain all the biological and clinical variables collected from samples during the project. It has advanced features such as data verifications, audit logs, modification of CRFs, query and discrepancy management, data extraction, etc. At the same time, this module will allow not only the collection, storage and management of clinical data in an electronic and secure way, but also the data export in several formats for subsequent analysis and exploitation by HUTER researchers. Furthermore, this component includes a user role system that is based on the roles typically found in clinical trials and other forms of clinical research to manage the authorized actions of each user.

This application has been improved in two development steps. First version included a Keycloak adapter to be integrated with HUTER Platform IAM. It also provided some features requested by project coordinators in

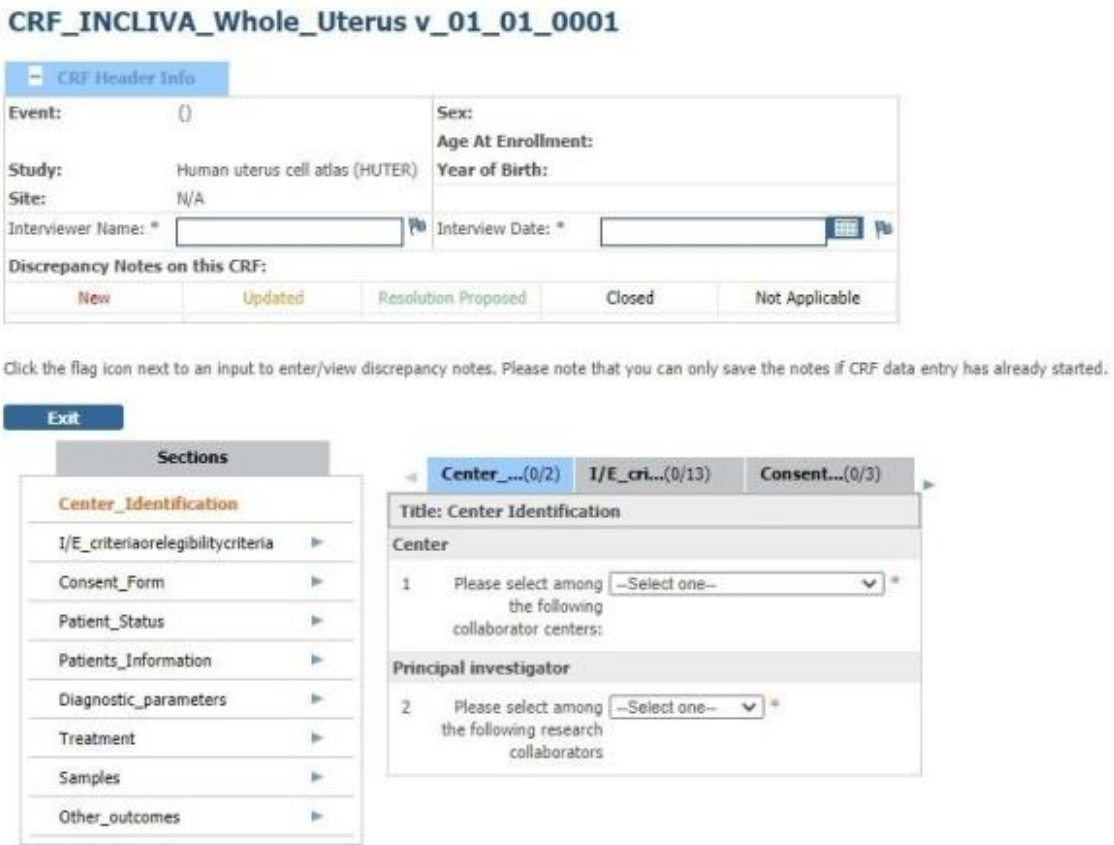
order to facilitate the generation of subject identifiers. Next pictures (from Picture 1 to 5) show the aspect of LibreClinica v.01.01 which includes just functional developments.



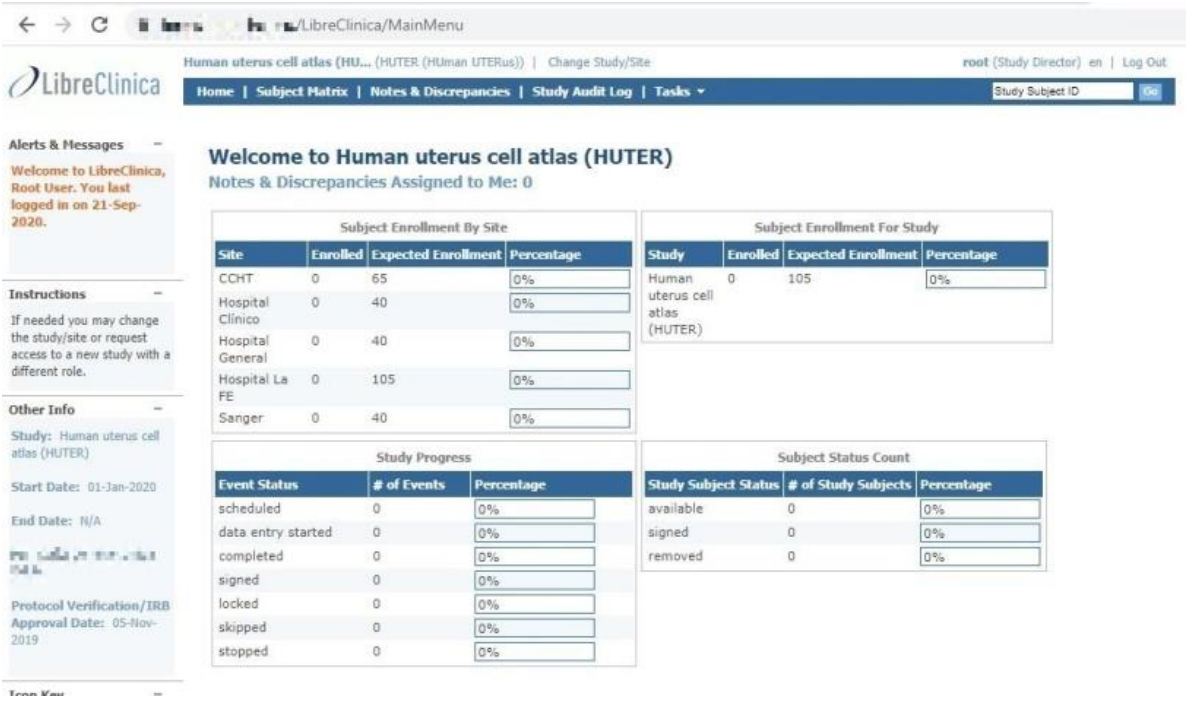
Picture 1 - HUTER specific generator for subject identifier



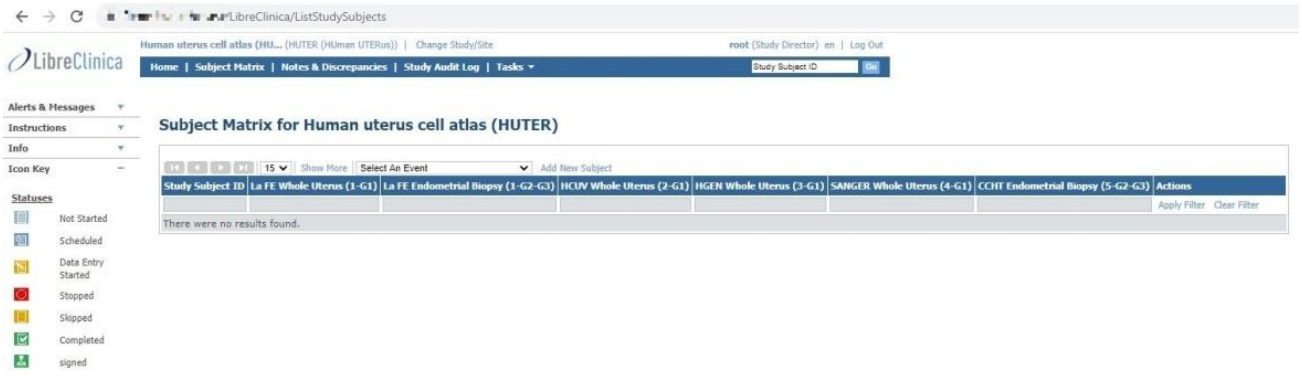
Picture 2 - HUTER study sites organization



Picture 3 - LibreClinica CRF data introduction page and early version of new navigation menu

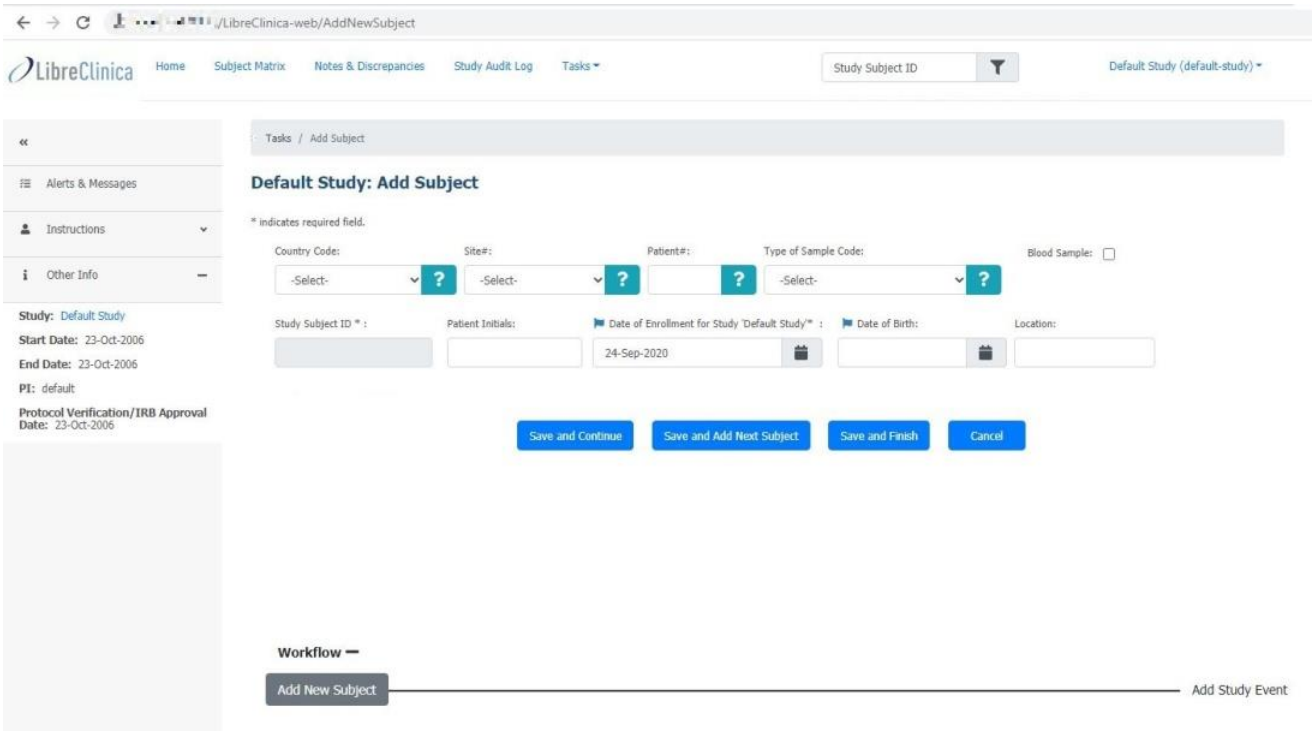


Picture 4 - LibreClinica original layout

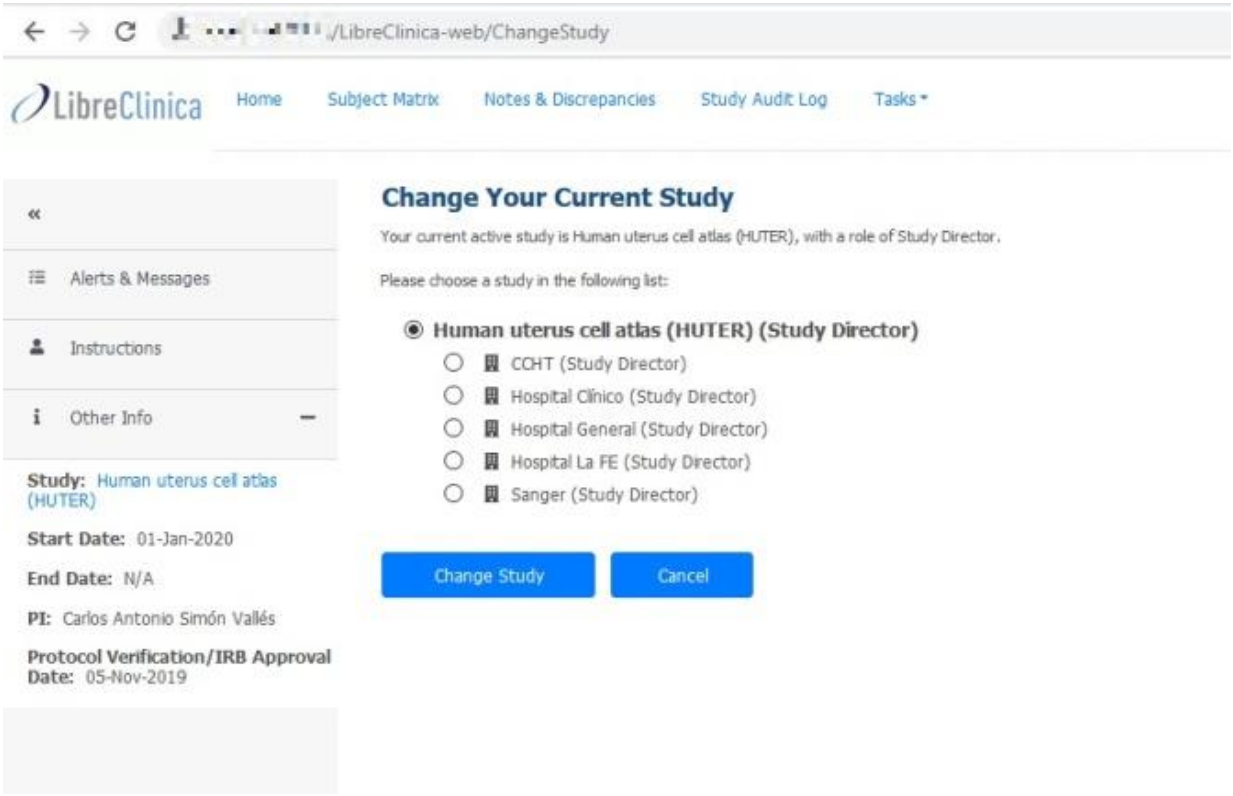


Picture 5 - LibreClinica original Subject Matrix

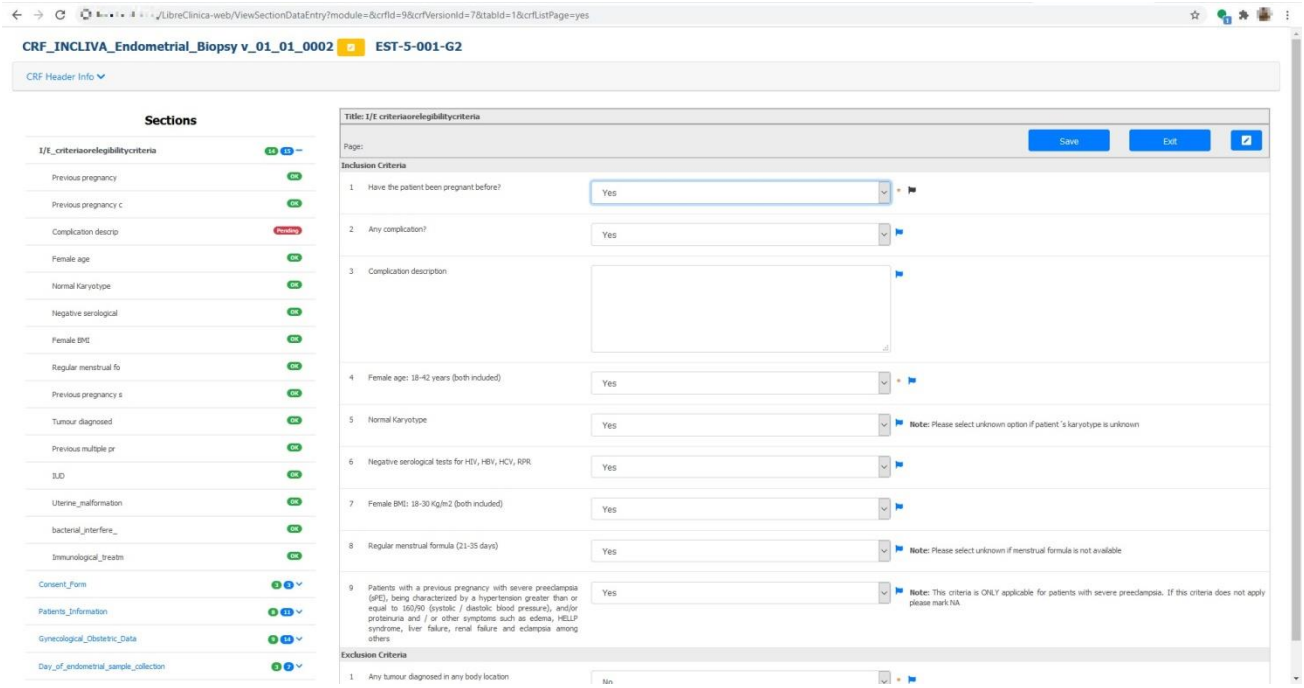
In order to provide a better user experience and add usability to LibreClinica, version v.01.02 includes new aspect that allows LibreClinica to get advantage of current display dimensions. Furthermore, some visual functionalities have been integrated to reach a more intuitive interface (see Pictures from 6 to 10).



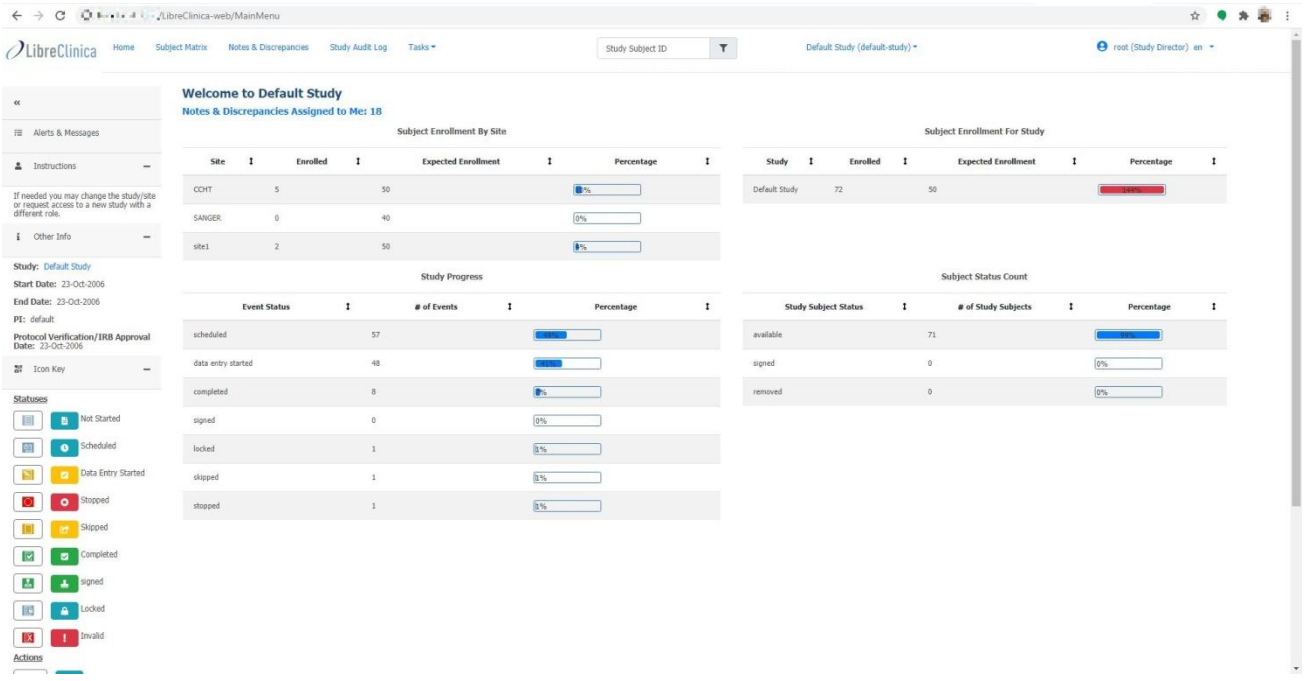
Picture 6 - LibreaClinica v.01.02 subject registration page



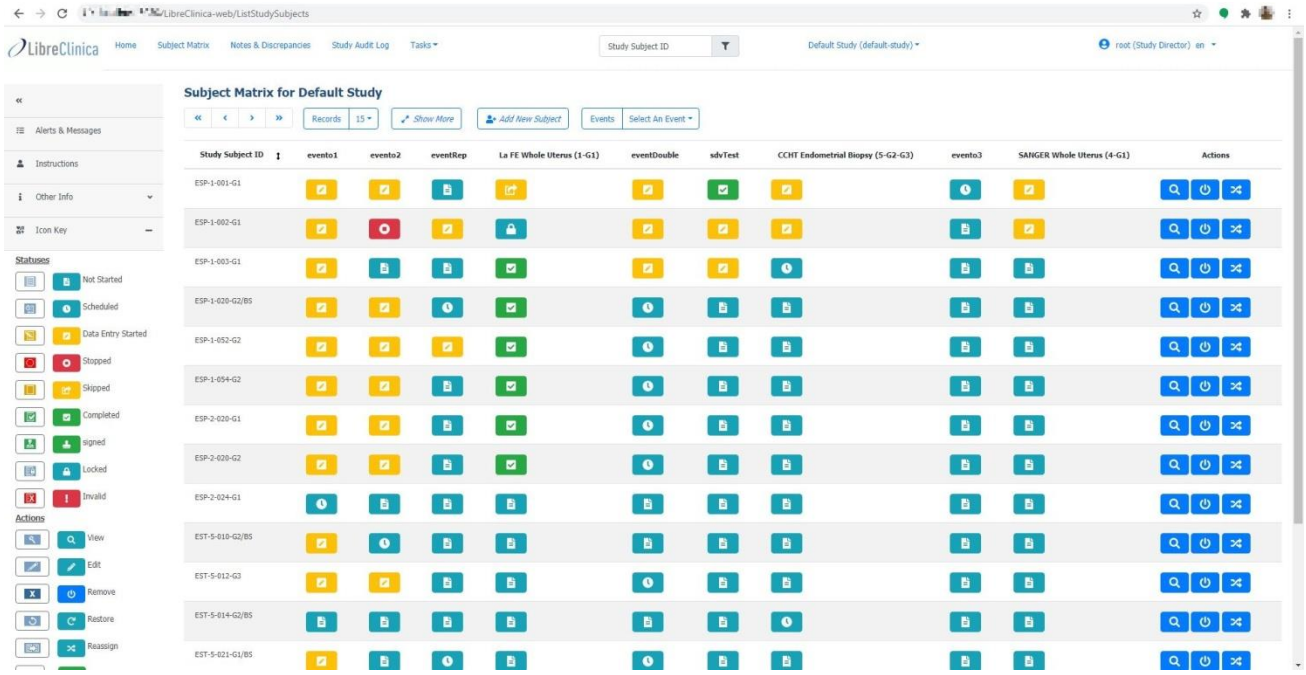
Picture 7 - LibreClinica v.01.02 List of Study Sites



Picture 8 - LibreClinica v.01.02 Data introduction page and new navigation menu



Picture 9 - LibreClinica v.01.02 layout

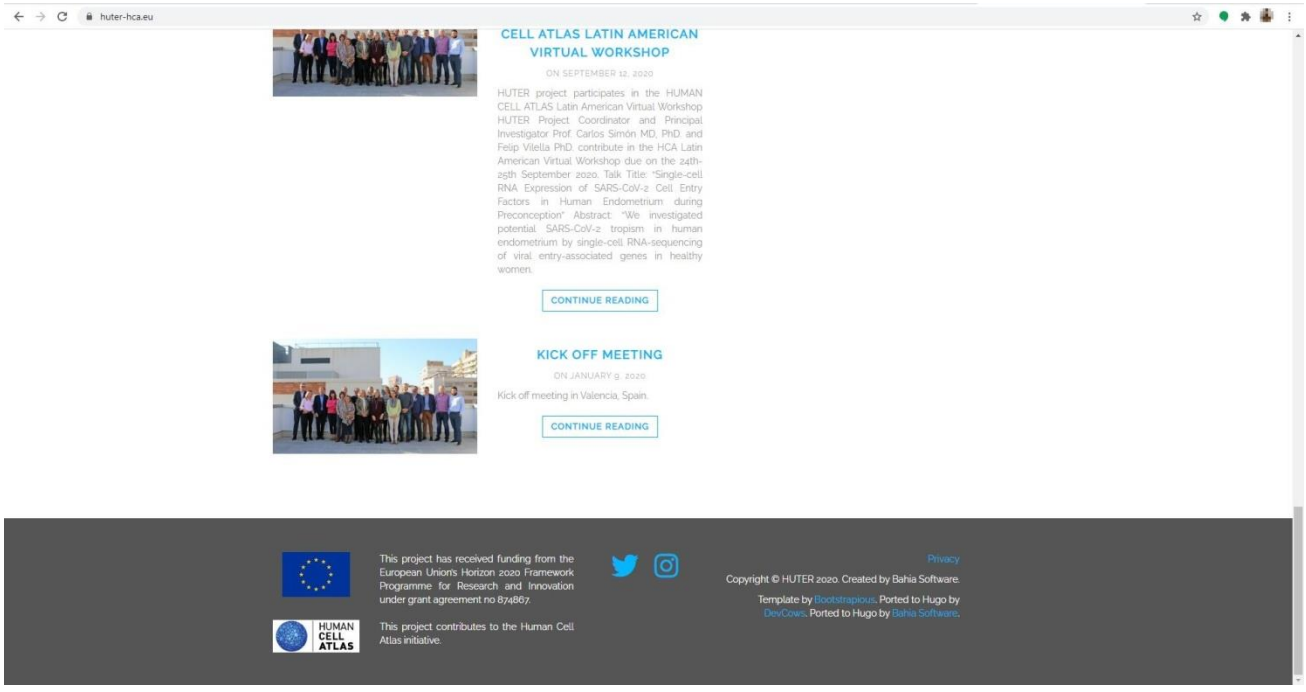


Picture 10 - LibreClinica v.01.02 Subject Matriz layout

4.3. Website

This component contains the public HUTER official website (https://huter-hca.eu/). It will become in the core of digital communication and dissemination activities of research results throughout the lifetime of the HUTER project. It will also contain a point of access to the different private modules through securely authentication with exclusive user credentials for HUTER users. The public website portal will not host personal or sensible data and it is totally public. See related pictures from 11 to 14.

Picture 11 - HUTER website



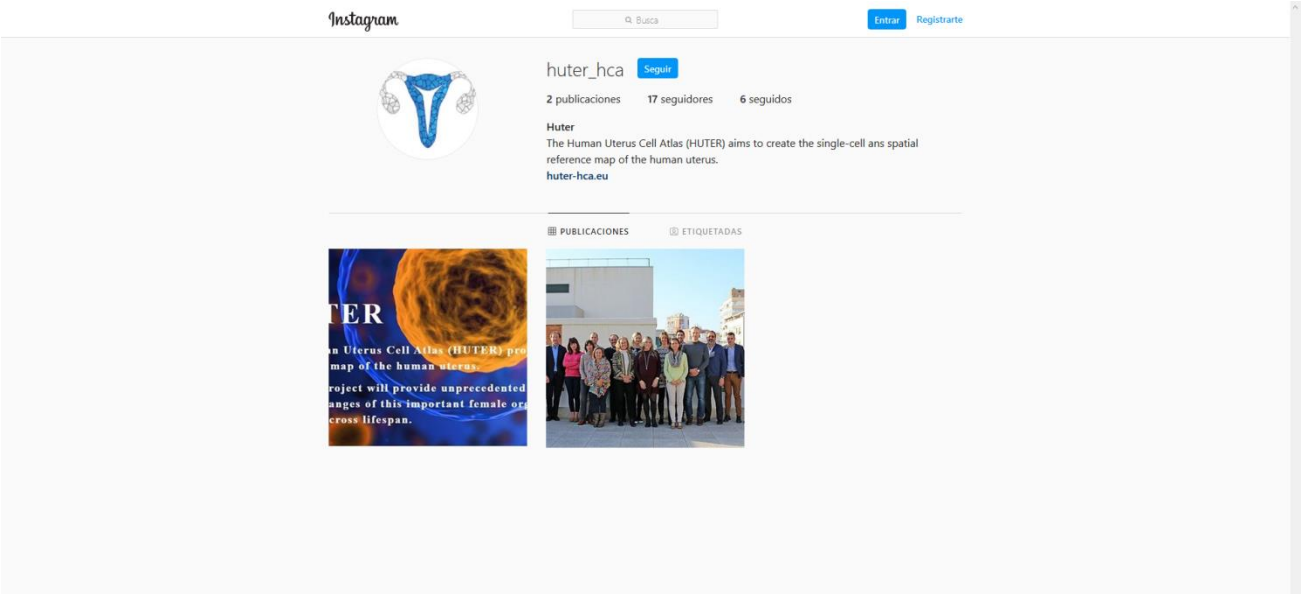
Picture 12 - HUTER website european project reference

Picture 13 - HUTER website news section

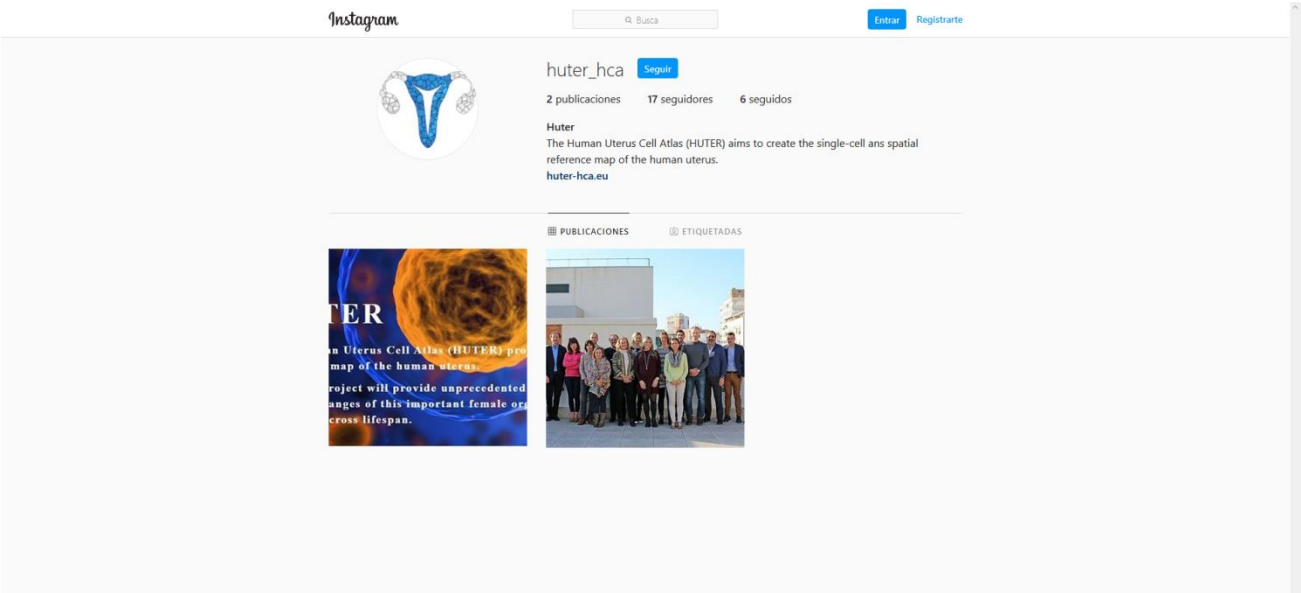
Picture 14 - HUTER website partners' section

4.4. Social network

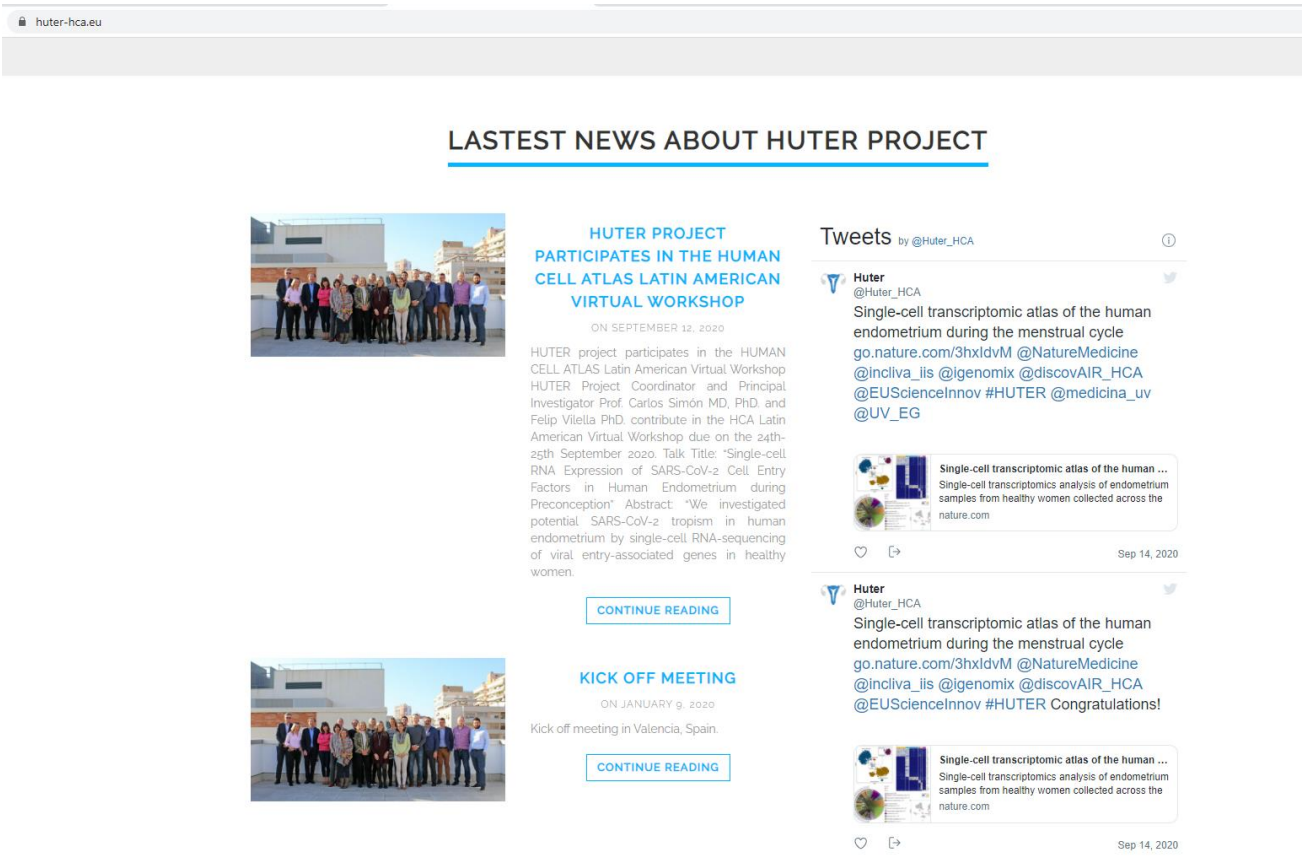
Even though this is not a component deployed under the HUTER Platform, social media accounts are required as tool for HUTER dissemination aim. Accounts in two well-known social networks, Twitter and Instagram were created with the aim of improve the communication activities related to the project. Our website infrastructure is ready to support not only links to these social networks, but also specific widgets to show the last communications of HUTER on these networks. In addition, this component is open to incorporate new social networks if the project requires it (see related pictures from 15 to 17).



Picture 15 - HUTER Twitter



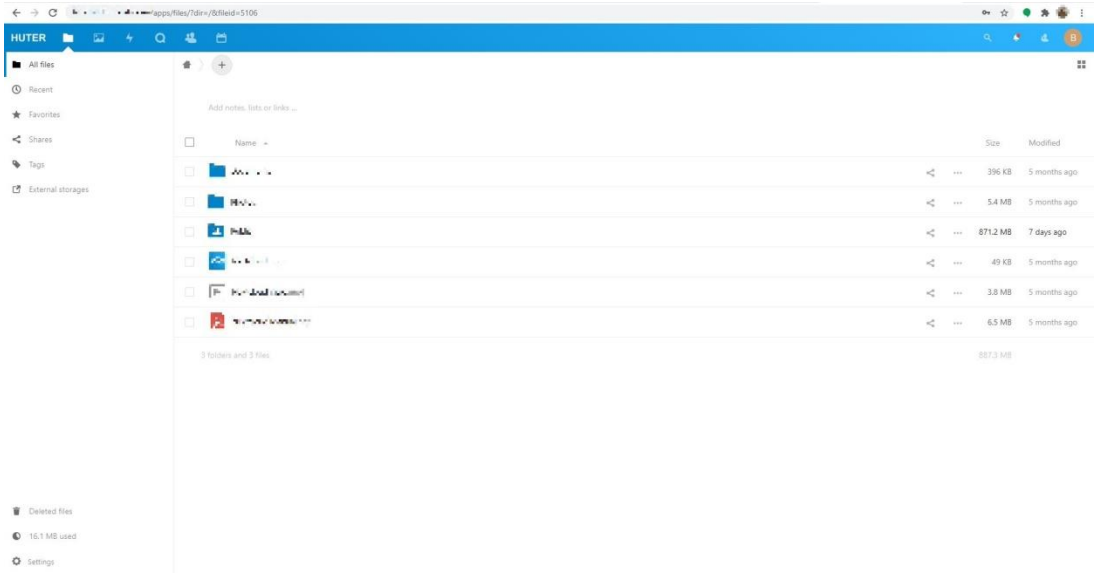
Picture 16 - HUTER Instagram



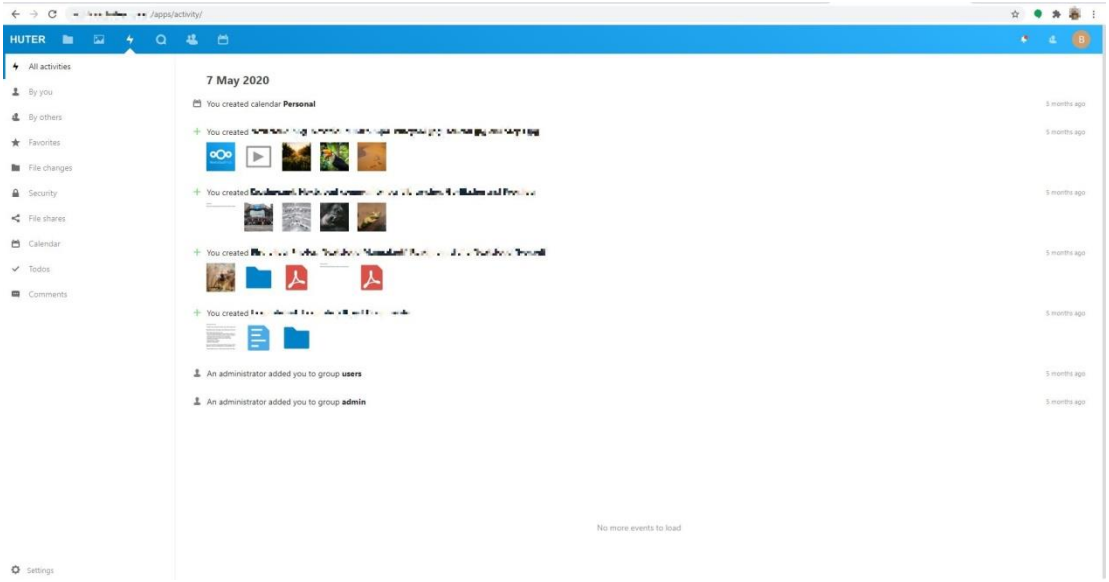
Picture 17 - Twitter feed in HUTER website

4.5. NextCloud + Plugins

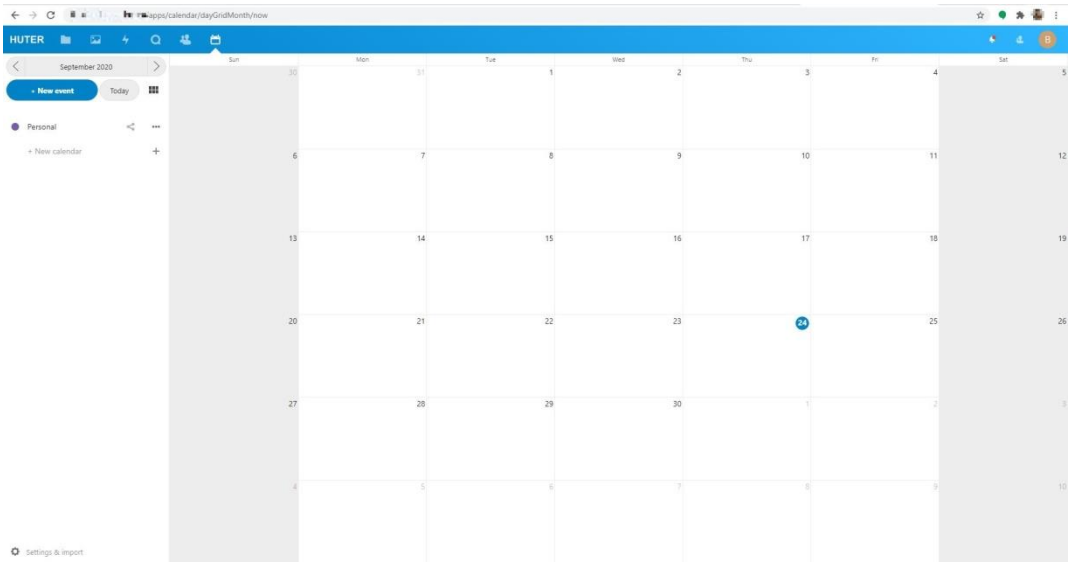
NextCloud and their plugins constitute the "intranet" component. It aims to improve the HUTER project managing tasks between partners located in different countries. For this purpose, NextCloud gathers a useful suit of open source tools, designed with a user-friendly interface for collaborative work, data sharing, videoconference, shared calendars, office tools, etc. This solution offers low adaptation period to learn to use this platform because the similarity with other popular private tools, such as Google Drive, Dropbox or Microsoft OneDrive. The data and information gathered in this module will be related to coordination tasks and documents repository (see related Pictures from 18 to 24).



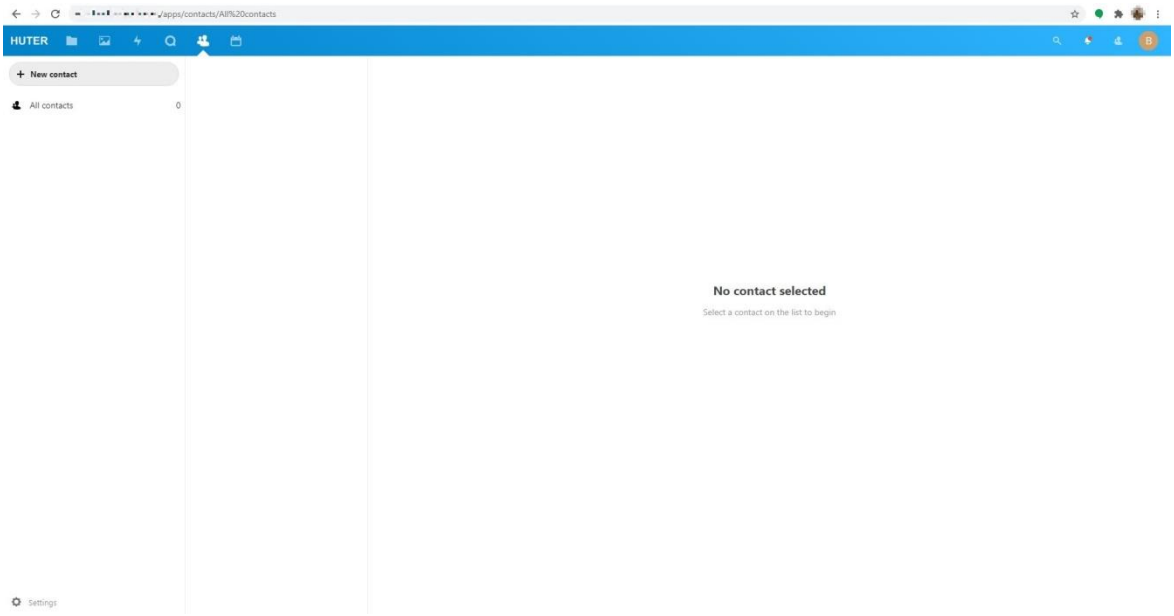
Picture 18 - Nextcloud sharing space



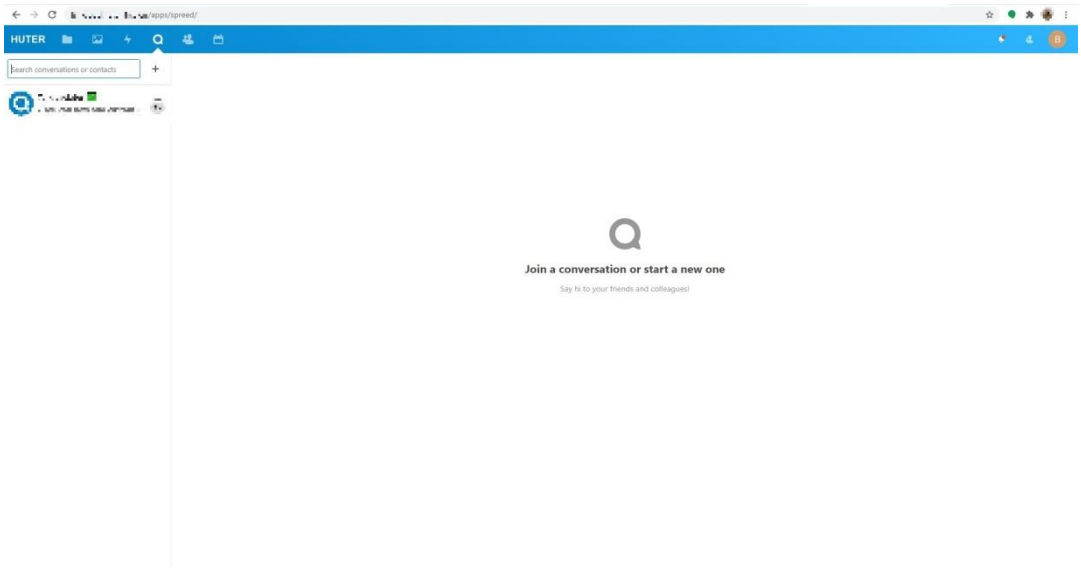
Picture 19 - Nextcloud activity dashboard



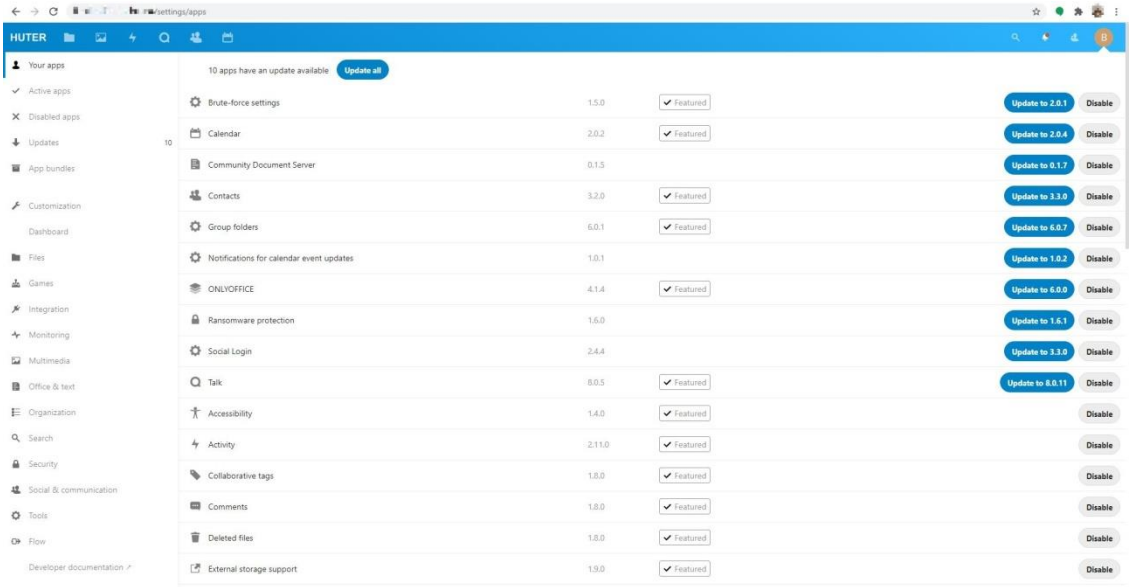
Picture 20 - Nextcloud calendar



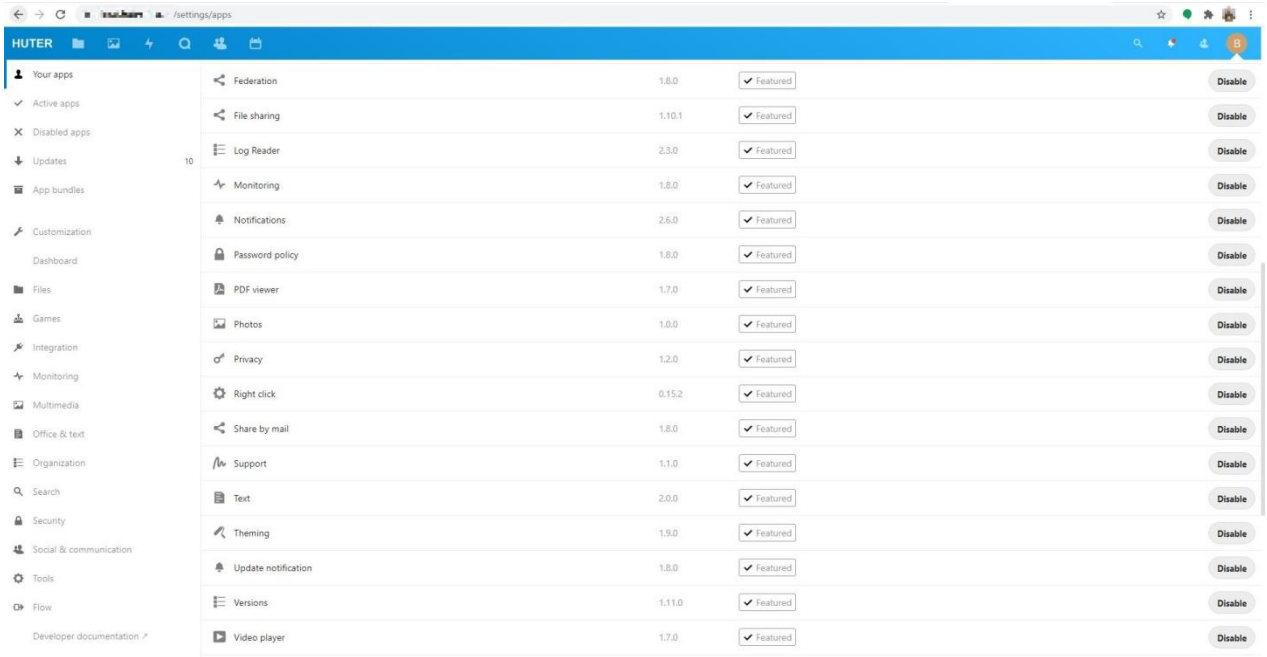
Picture 21 - Nextcloud contact list



Picture 22 - Nextcloud meeting app



Picture 23 - Nextcloud plugin list 1

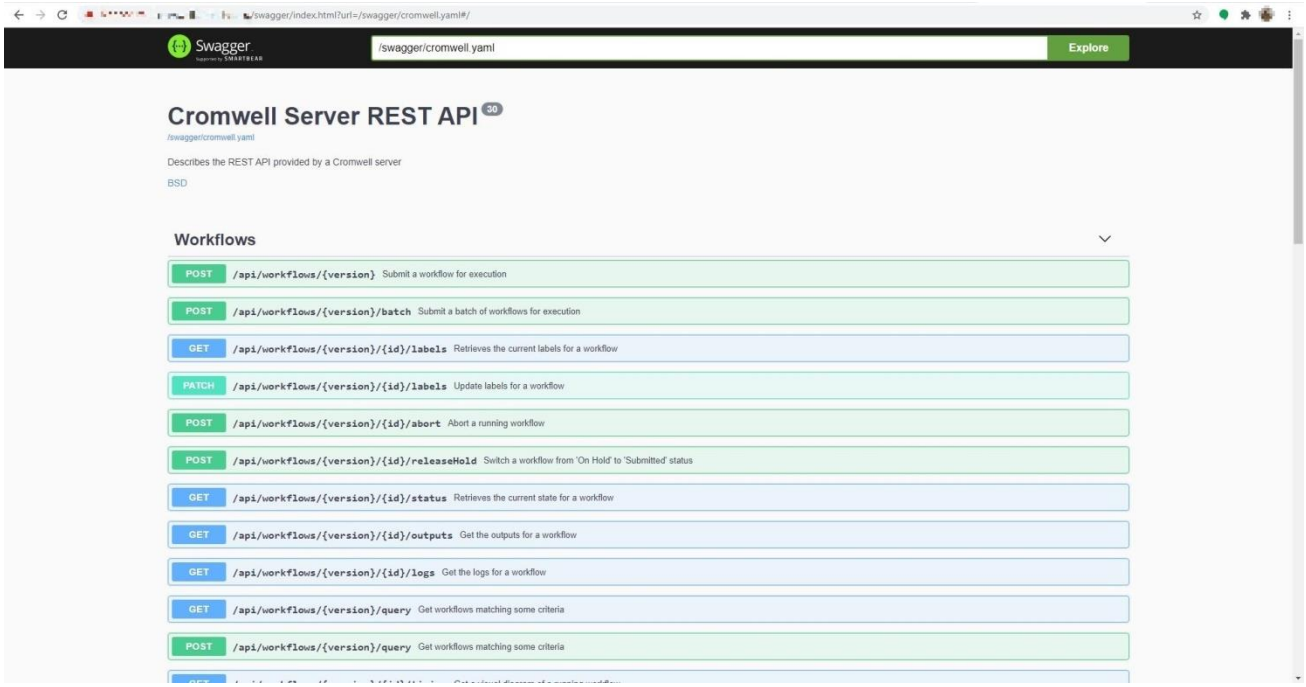


Picture 24 - Nextcloud plugins list 2

4.6. Pipeline Manager (Cromwell) + Scalable processing backend (AWS Batch)

HUTER Platform contains an area to execute pipelines over cell sequencing data in order to analyze them or generate new results in form of new data files. HUTER platform is aligned with HCA-DCP in this regard deploying a Pipeline Manager based on Cromwell. Cromwell is one of the workflow management systems recommended by HCA that allows scientists to run pipelines written in WDL (Workflow Description Language) over data. Additionally, a scalable processing backend is enabled to handle these executions. In this case, AWS batch service was configured to dynamically provision the optimal quantity and type of computing resources (e.g., CPU or memory optimized instances) based on the volume and specific resource requirements of the batch jobs submitted.

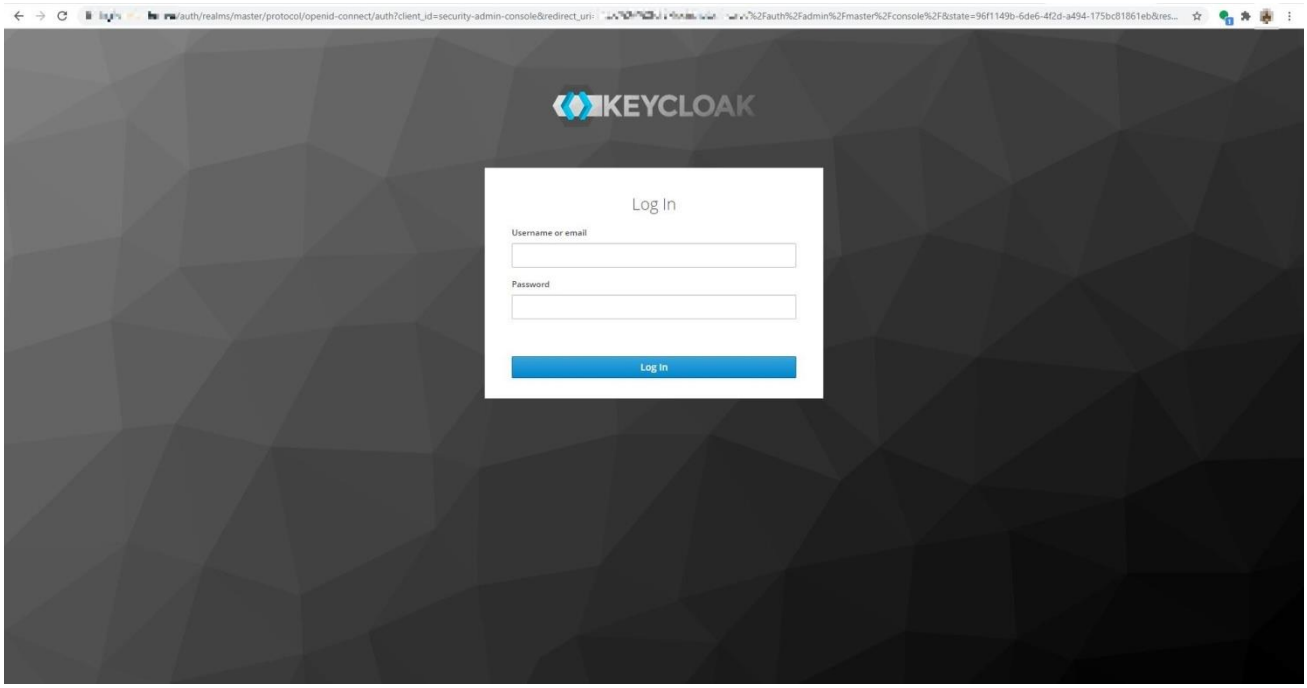
Pipeline execution infrastructure is already deployed and working integrated with a pure AWS S3 storage. Thus, WDL can be run over training data in order to ensure analysis process workflow. The API of the deployed Cromwell is shown as evidence in Swagger format, so test executions can be done from any website. However, additional AWS S3 access is required in order to provide input data. In regards of WDL, even when git repositories are ready to store WDL files, they can be provided to Cromwell through the Swagger interface (see Picture 25).



Picture 25 - Cromwell API in Swagger format

4.7. Identity and access management component

The HUTER platform will use KeyCloak software as advanced open source identity and access management system. It allows not only to configure permissions for each component and user profile but also to create an access record that ensures the accesses traceability (see Picture 26).

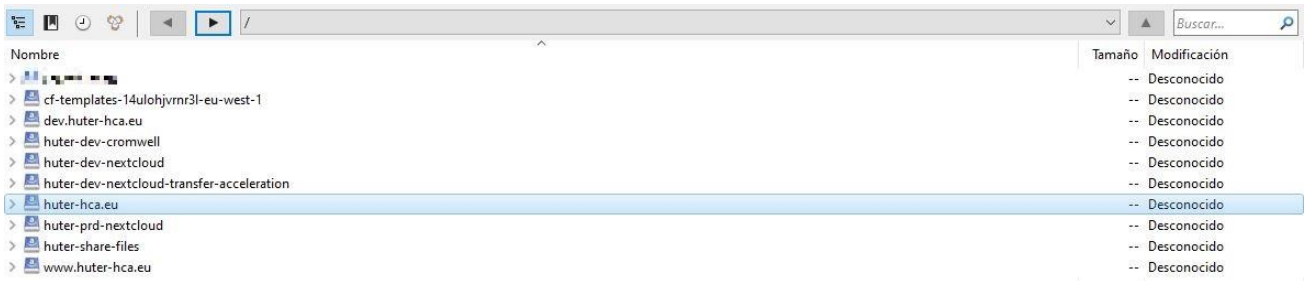


Picture 26 - Keycloak for HUTER SSO login page

4.8. S3 (infrastructure)

Amazon Simple Storage Service (S3) is an object storage service that offers industry-leading scalability, data availability and security. The infrastructure required to support S3 is ready and deployed to be used by HUTER platform. This means that HUTER platform can use it to store and protect any amount of data. Furthermore, it provides easy-to-use management features so data can be configured with finely-tuned access controls to meet HUTER specific requirements. S3 automatically creates and stores copies of all S3 objects across multiple systems so that data is available when needed and protected against failures, errors, and threats.

Some AWS S3 buckets have been created to accomplish specific functionalities like website storage or Cromwell related space for input and output files (see Picture 27).



Picture 27 - AWS S3 current buckets

4.9. DICOM PACS

A picture archiving and communication system (PACS) is a medical imaging technology which provides economical storage and convenient access to images from multiple sources. This component is deployed over the HUTER infrastructure with the aim of storing and managing the DICOM images that comes from different advanced research microscopes of HUTER researchers (see Picture 28).

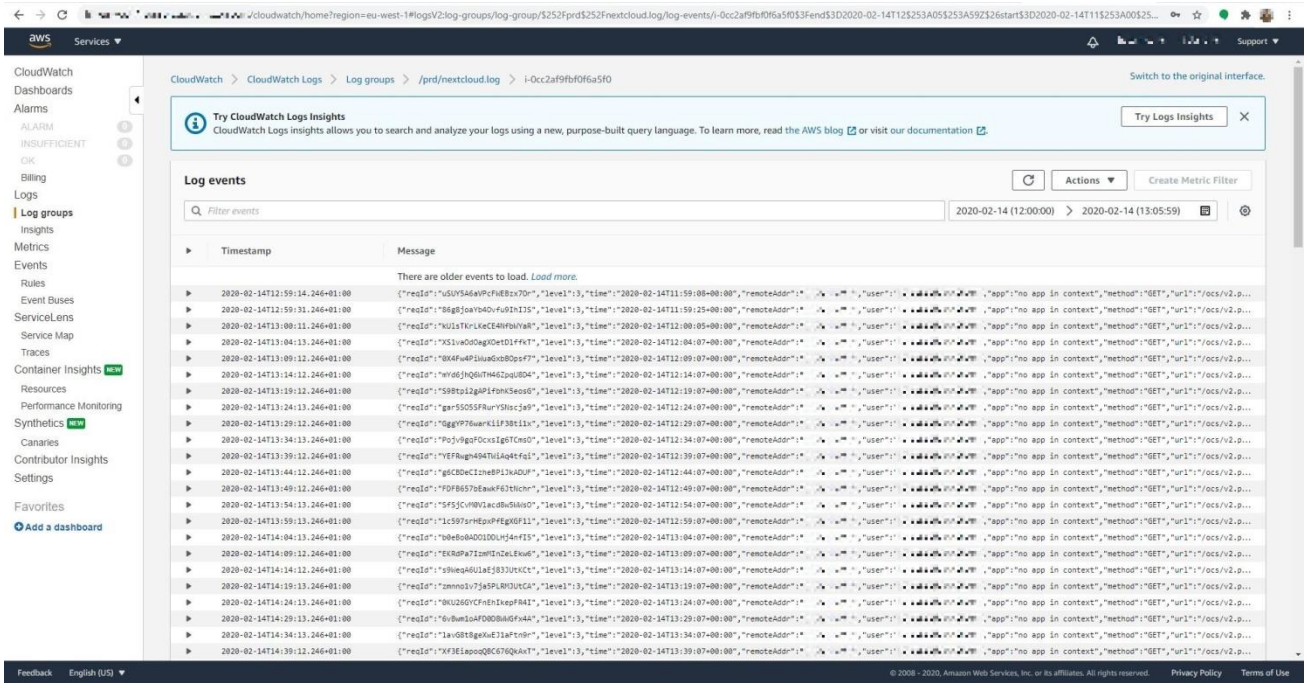
Picture 28 - dcm4chee Archive 5 web portal

4.10. DICOM viewer (infrastructure)

A custom infrastructure to support the state-of-art DICOM viewer compatible with advanced images to be generated by the latest single cell sequencing and microscopy techniques used by HUTER project researchers has been deployed. However, the DICOM viewer is currently under development and the final version is expected to be provided as part of Deliverable D2.4. (HUTER_WP2_D2.4_Final_implementation_of_data_access_tools_and_DICOM_visualization_tool) during month 21 of the project.

4.11. Monitoring & Log Management

CloudWatch is a service provided by AWS to collect monitoring and operational data in the form of logs, metrics, and events of the AWS' services and modules installed in HUTER platform. This component has been deployed and it will provide security and traceability of user actions and data in the platform (see Picture 29).



Picture 29 - Amazon Cloudwatch for HUTER Platform

5. NON-DEPLOYED COMPONENTS

According to the Project Plan, some tools are not deployed yet because they are in an early development stage or they are in a designing process. However, the architecture to support these tools has been fully defined in Deliverable 2.1 and it is ready to support all these tools when their development process ends. In this section these tools will be listed in order to clarify dimension of the complete HUTER Platform.

Components	Description
LibreClinica webservice layer	Service that allows the integration with LibreClinica registered information.
Ingestion manager	Service layer for submission managing. It allows broker to request access to the HUTER storage.
Broker-uploader	Client application for file uploading to HUTER Platform. Ensures data integrity and subject tracking.
Storage Manager	Service layer for file consolidation in the HUTER Platform.
Git repository	Storage for WDL and dockerfiles for pipeline definition and execution.
Docker repository	Docker repository for pipelines execution.

	Autobuild from dockerfiles.
Pipeline Web Manager	Web portal to manage Cromwell execution requests integration S3 storage, git WDL, metadata and Cromwell.
Dicom Viewer	The purpose of this viewer is to allow researchers to examine uterus tissues and cells with very high resolution from these advanced images in their laptops. It will allow not only the rapid identification of areas of interest but also collaborative studies of very specific cell shapes and arrangements that are usually associated with the disease. Furthermore, the DICOM viewer will have the capability to integrate data from very complex analysis that HUTER scientists will carry out with the proprietary software installed in their imaging equipment and to provide basic features for Virtual Microscopy approaches such as identification of abnormalities, addition of comments, image segmentation, etc. Additionally, this component will enable synchronisation with corresponding PACS component. These features will ensure an optimal user experience during access, visualization and manipulation of the digitized samples, simulating as far as possible the interaction with current visualization tools broadly used in research, thus the system will be intuitive, and the learning curve will be low.
Data Web Browser	Web application for data querying in the HUTER Platform. It provides the capability of download index files for raw data download.
Downloader	Client application for raw-data downloading from HUTER Platform to a device by an authorized user.
HCA Mapping Process	Support application for HCA metadata building based on the HUTER stores data.

Intranet access	Private area under the website for registered users in HUTER platform. It allows users to have a personal view of the platform and facilitates their access to application according to their profile.
Configuration Server	Service for managing application variables per environment. It offers a centralized point for every application an environment under the HUTER Platform. It is also needed for microservices scalability.
Registration and Discovery Server	These components provide a unique point to know microservices running instances. In works synchronously to Gateway and configuration server in order to get microservices balancing and scalability.
Microservices Gateway	Service that provides service load balancing and is integrated with Discovery service to query available service instances.