Progettazione piattaforma ISTI per il progetto SerGenCovid-19

Franca Debole[1]*, Andrea Dell'Amico[1], Tommaso Piccioli[1], Enrico Fantini[1], Giuseppe Lipari[1], Federico Volpini[1]

Sommario
Nel contesto del progetto di ricerca denominato "SerGenCovid-19 (Serum Genetic Covid-19 study) Indagine sierologica e genetica sull'immunità e la suscettibilità all'infezione da SARS-CoV-2 e creazione di una biobanca", l'ISTI è coinvolto come responsabile nel *Work Package 6: Progettazione e implementazione della piattaforma informatica per la gestione di "Raccolta, conservazione e consultazione dei dati sanitari relativi ai prelievi ematici"*. In questo report tecnico vengono descritte le scelte di progettazione della suddetta piattaforma informatica per conservare e consultare i dati clinici anonimizzati dei partecipanti al progetto.

Keywords
SARS-CoV-2 — Covid-19 — Questionari Anmanestici — Piattaforma Web

[1] *Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy*
***Corresponding author**: franca.debole@isti.cnr.it

Indice

Introduzione

Il Dipartimento di Scienze Biomediche, di seguito denominato "DSB" ha intrapreso un nuovo progetto di ricerca denominato "*SerGenCovid-19 (Serum Genetic Covid-19 study) Indagine sierologica e genetica sull'immunità e la suscettibilità all'infezione da SARS-CoV-2 e creazione di una biobanca*" per individuare la tipologia di anticorpi neutralizzanti e di mediatori immunologici solubili nel tempo, studiare l'influenza della genetica nel determinare la qualità di risposta all'infezione da SARS-CoV-2 mediante l'analisi dello stesso campione e di studiare il rapporto tra sieroprevalenza, biomarcatori e condizioni ambientali che includono i fattori climatici, inquinanti atmosferici, tipo di residenza rurale o urbana dei partecipanti. Il progetto, vede coinvolti i seguenti istituti:

- l'Istituto di Fisiologia Clinica (IFC), l'Istituto per l'Endocrinologia e Oncologia Sperimentale (IEOS),
- l'Istituto di Genetica Molecolare "Luigi Luca Cavalli-Sforza" (IGM),
- l'Istituto di Ricerca Genetica e Biomedica (IRGB),
- l'Istituto per la Ricerca e l'Innovazione Biomedica (IRIB),
- l'Istituto di Biologia e Patologia Molecolari (IBPM),
- il Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca, CNR,
- l'Istituto di Informatica e Telematica Sede di Pisa (IIT),
- l'Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" (ISTI).

Lo studio proposto si articola in quattro obiettivi principali:

1. valutare le modificazioni della risposta immunitaria alla malattia o al vaccino a distanza di tempo;
2. valutare l'effetto delle procedure vaccinali sull'andamento temporale della diffusione virale;
3. studiare il rapporto tra sieropositività, biomarcatori e condizioni ambientali;
4. la nascita di una biobanca con i campioni biologici, finalizzata a rendere possibili studi ulteriori di associazione genetica e di identificazione di biomarcatori prognostici e di suscettibilità.

Per poter raggiungere i suddetti obiettivi, SerGenCovid-19 prevede una raccolta di dati clinici, sieri e materiale genetico su larga scala nella popolazione italiana. A partire da 100.000 soggetti reclutati nell'ambito dello studio EPICOVID-19 nella primavera del 2020, che hanno manifestato la disponibilità a essere ricontattati per ulteriori studi, verranno selezionati su base volontaria 10.000 partecipanti, distribuiti omogeneamente in tutto il paese. I volontari che avranno dato il loro

consenso alla partecipazione al progetto saranno sottoposti a 3 test seriologici (T0, T1 e T2 a distanza di 5 mesi l'uno dall'altro) per l'analisi dei livelli anticorpali anti SARS-CoV-2. In aggiunta, dato che alcuni studi evidenziano l'ipotesi che la suscettibilità all'infezione e l'andamento clinico della malattia nelle sue diverse forme possano essere influenzati anche da fattori genetici, per tale motivo ai volontari verrè prelevato un campione di DNA (al tempo T0), allo scopo di effettuare studi di associazione genetica genome-wide e/o analisi di specifici genotipi eventualmente correlati all'infezione. Per portare a termine i 4 obiettivi sopraelencati il progetto si articola in sei work package:

- WP1. Studio sierologico
- WP2. Studio genetico
- WP3. Analisi epidemiologica
- WP4. Creazione della biobanca
- WP5. Progettazione e implementazione della piattaforma informatica per la gestione di "Informativa, consenso, e raccolta dati anagrafici"
- WP6. Progettazione e implementazione della piattaforma informatica per la gestione di "Raccolta, conservazione e consultazione dei dati sanitari relativi ai prelievi ematici".

Mentre i WPs da 1-4 riguardano le problematiche più strettamente correlate alla parte biomedica del progetto, quindi la raccolta dei campioni ematici e di DNA e gli studi a essi correlati, il WP5 e il WP6 riguardano gli aspetti piu tecnici informatici per la corretta realizzazione di questa indagine, che spaziano dalla selezione e le comunicazioni con i partecipanti fino alla memorizzazione, conservazione e consultazione dei risultati dei vari test effettuati. In particolare, l'ISTI è stato coinvolto come responsabile nel *Work Package 6: Progettazione e implementazione della piattaforma informatica per la gestione di "Raccolta, conservazione e consultazione dei dati sanitari relativi ai prelievi ematici"*.

1. Analisi dei requisiti

ISTI è responsabile della progettazione e realizzazione di una piattaforma informatica per conservare e consultare i dati clinici anonimizzati dei partecipanti a questa iniziativa. In particolare, la piattaforma deve fornire accesso a tre tipologie di utenti diversi, con funzionalità e interfacce diverse. Da una prima analisi dei requisiti, conclusasi a Aprile 2021, la piattaforma ISTI, prevede i seguenti utenti/utilizzatori:

- il partecipante alla campagna;
- l'operatore che inserisce i risultati dei test seriologici e dei test del DNA;
- il ricercatore che consulta i dati.

Per ognuno dei suddetti utenti, ISTI sarà impegnato nella progettazione e nello sviluppo di una piattaforma informatica che fornirà tre diverse applicazioni:

- un'interfaccia web dedicata ai partecipanti per la compilazione del questionario anamnestico elaborato dagli esperti;
- un'interfaccia web per la compilazione delle schede per i risultati dei test seriologici dedicata agli operatori dei laboratori designati;
- un'interfaccia web per la consultazione dei dati raccolti dedicata ai ricercatori designati

1.1 Le interfacce web

Interfaccia Partecipante. Il partecipante alla campagna tramite accesso alla piattaforma realizzata dallo IIT, potrà:

- avere accesso alla compilazione del questionario anamnestico;
- avere accesso per la sola consultazione al questionario compilato solo in modalità lettura;
- accedere ai referti dei test seriologici in formato pdf.

Interfaccia Operatore. L'operatore addetto all'inserimento dei risultati dei test seriologici avrà accesso al portale apposito implementato come una VRE in D4Science [1] e potrà:

- inserire i referti dei test seriologici (3) per ogni partecipante;
- consultare i referti dei test seriologici (3) per ogni partecipante;

Interfaccia Ricercatore. Il ricercatore avrà accesso ai questionari e ai risultati anonimizzati tramite il portale apposito implementato come una VRE in D4Science e potrà:

- scaricare i dati memorizzati nel DB in formato CSV, disponibili periodicamente nel workspace della VRE.

2. Progettazione

2.1 Caratteristiche generali

In base alla specifica dei requisiti prodotta dall'analisi, nella progettazione sono state prese in considerazione anche le seguenti caratteristiche fondamentali:

- necessità di due server gestiti da un load balancer per bilanciare il carico delle richieste e per assicurare resistenza ai guasti;
- necessità di un database con supporto alla cifratura delle informazioni memorizzate;
- necessità di effettuare un backup dei dati.

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

2.2 Gli accessi

Per le tre tipologie di utenti (partecipante, operatore, ricercatore) delle tre interfacce si prevedono diverse tipologie di accesso:

- il partecipante si collegherà all'interfaccia ISTI tramite il portale realizzato dallo IIT usando le credenziali gestite da IIT;
- l'operatore potrà accedere dal portale D4Science usando le credenziali che preferisce tra quelle supportate (Accademiche, Google, Twitter eccetera) e previa autorizzazione;
- il ricercatore potrà accedere dal portale D4Science usando le credenziali che preferisce tra quelle supportate (Accademiche, Google, Twitter eccetera) e previa autorizzazione.

2.3 I dati

Ogni dato clinico (prelievi ematici, questionario anamnestico) deve essere codificato senza i dati identificativi del partecipante. Più precisamente, le due piattaforme, quella in carico allo IIT e realizzata nel WP5 e quella in carico a ISTI e realizzata nel WP6 devono essere completamente separate e mentre la prima conterrà solo le informazioni identificative del partecipante (come nome, cognome, email eccetera), la piattaforma ISTI conterrà solo i dati clinici del partecipante senza alcun dato identificativo esplicito. Per poter realizzare questa scissione delle due piattaforme, mantenendo la separazione dei dati identificativi dai dati clinici, al partecipante verrà associato un codice univoco (cua) assegnato dalla piattaforma IIT.
I dati contenuti e gestiti nella piattaforma ISTI sono:

- il codice univoco assegnato dalla piattaforma del WP5 a coloro che avranno aderito all'iniziativa;
- i dati raccolti con il questionario anamnestico opportunamente criptati e pseudonimizzati;
- i risultati delle analisi immunologiche e sierologiche opportunamente criptati e pseudonimizzati.

Tutti questi dati saranno riversati in un DBMS e saranno successivamente resi disponibili ai ricercatori. Per la sicurezza dei dati, questi saranno memorizzati in un DBMS, residente su uno o più server di un'infrastruttura informatica costantemente monitorata usando meccanismi quali firewall, sistemi di monitoraggio degli accessi e log delle attività e inoltre per una maggior protezione i dati della piattaforma verranno cifrati.

3. Tecnologie

Così come mostrato in Figura 1 l'architettura della piattaforma è la seguente:

- load balancer già disponibile nell'infrastruttura dell'istituto;
- due VM con Ubuntu LTS per le applicazioni web;
- due VM con Ubuntu LTS per la base di dati configurate in replica sincrona.

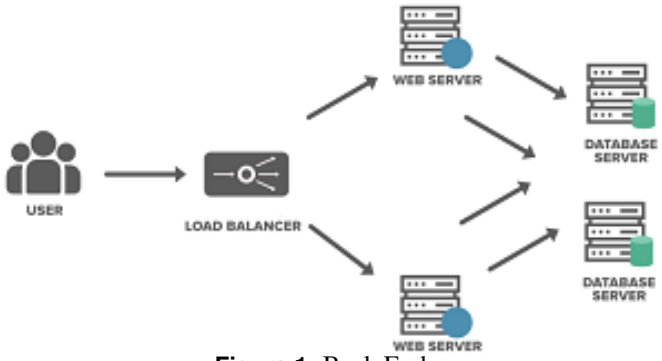


Figura 1. Back End

3.1 Database

Dall'analisi dei requisiti, sono scaturiti i seguenti dati da memorizzare nel database ISTI:

- i risultati anonimizzati dei 3 test sierologici;
- i risultati anonimizzati dei test genetici;
- i dati anonimizzati del questionario di anamnesi compilato dai partecipanti all'iniziativa;
- i dati del questionario Epicovid [formato probabile EXCEL];
- il codice questionario Epicovid.

Data la necessità di avere dati cifrati la scelta del sistema di basi di dati è ricaduta sul database PostgreSQL (versione 13) che ha appunto un'estensione pg_crypto pensata per la crittografia dei dati.

In Figura 2, è esposto lo schema della base di dati per memorizzare le informazioni necessarie, mentre in Tabella 1 ci sono i dettagli delle tabelle del DB:

- **partecipante:** tabella contenente i dati associati ai partecipanti;
- **sierologico_test:** tabella contenente i dati dei test sierologici;
- **dna_test:** tabella contenente i dati dei test genetici;
- **sgc19_quest:** tabella contenente i questionari anamnestici;
- **epicovid_quest:** tabella contenente i questionari forniti dal progetto Epicovid.

3.2 Interfaccia Web

Nell'affrontare la progettazione delle tre interfacce abbiamo costruito dei wireframe per le tre tipologie di utenti così come illustrato nella Figura 3.

Figura 2. Database Schema

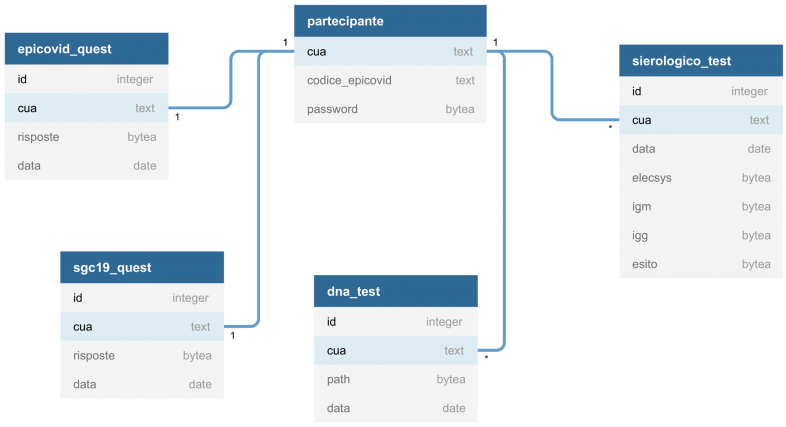


Tabella 1. Tabelle DB

partecipante

Attributo	Tipo	Descrizione
cua	UUID	codice univoco del partecipante
codice_epicovid	TEXT	il codice del questionario Epicovid
password	BYTEA	password per scaricare i sierologici

sierologico_test

Attributo	Tipo	Descrizione
id	INTEGER	l'indice univoco del test
cua	UUID	codice univoco del partecipante
data	DATE	la data di acquisizione
elecsys	BYTEA	il valore del Elecsys
igm	BYTEA	il valore del IgM
igg	BYTEA	il valore del IgG
esito	BYTEA	l'esito del test

dna_test

Attributo	Tipo	Descrizione
id	INTEGER	l'indice univoco del test
cua	UUID	codice univoco del partecipante
path	BYTEA	il percorso dei file contenenti i risultati dei test genetici
data	DATE	la data di acquisizione

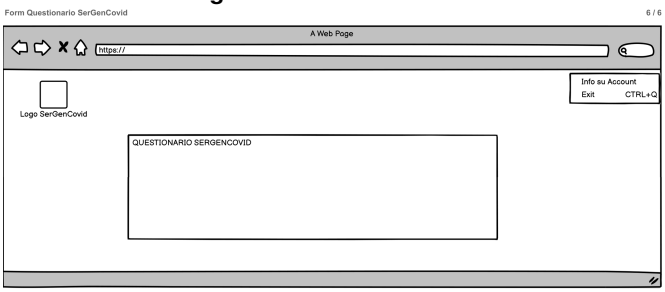
sgc19_quest

Attributo	Tipo	Descrizione
id	INTEGER	l'indice univoco del questionario
cua	UUID	codice univoco del partecipante
risposte	BYTEA	le risposte del questionario
data	DATE	la data di acquisizione

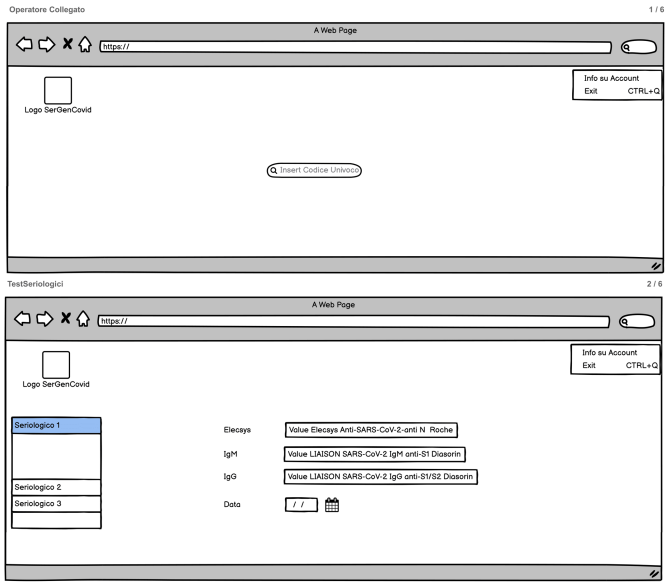
epicovid_quest

Attributo	Tipo	Descrizione
id	INTEGER	l'indice univoco del questionario
cua	UUID	codice univoco del partecipante
risposte	BYTEA	le risposte del questionario EPICOVID
data	DATE	la data di acquisizione

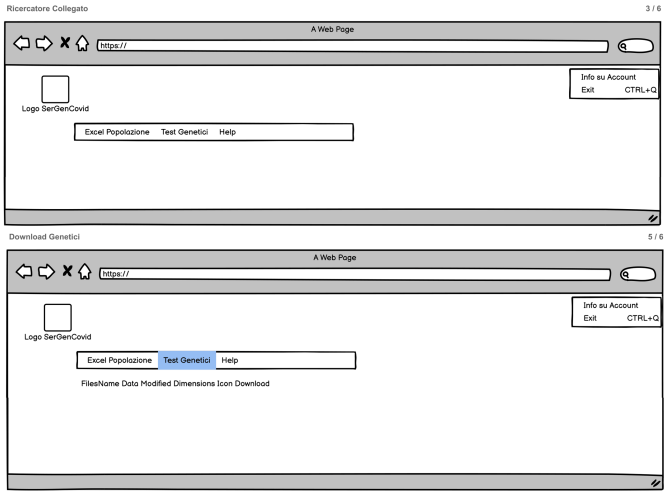
Figura 3. Wireframes UI.



(a) UI Utente.



(b) UI Operatore.



(c) UI Ricercatore.