# Realizzazione piattaforma ISTI per il progetto SerGenCovid-19

Franca Debole[1]*, Andrea Dell'Amico[1], Tommaso Piccioli[1], Federico Volpini[1], Giuseppe Lipari[1], Lorenzo Luconi Trombacchi[2], Maurizio Martinelli[2], Massimiliano Assante[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 rapporto di progetto viene descritta la prima fase del progetto dove vengono realizzate la parte di backend e le due piattaforme, una per il partecipante e una per l'operatore.

**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*
[2] *Istituto di Informatica e Telematica , 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.

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 questo contesto l'ISTI è responsabile del WP6 per progettare e implementare la piattaforma informatica per la gestione di "Raccolta, conservazione e consultazione dei dati sanitari relativi ai prelievi ematici". Dall'analisi dei requisiti questa piattaforma deve fornire accesso a tre tipologie di utenti diversi, con funzionalità e interfacce diverse.

Nei paragrafi seguenti vengono dettagliate la realizzazione delle tre componenti, necessarie nella prima fase del progetto: il backend, la piattaforma *partecipante* e la piattaforma *operatore*.

## 1. Descrizione Generale

Di seguito riportiamo i punti salienti dei requisiti, così come stati definiti in fase di analisi e progettazione, considerati

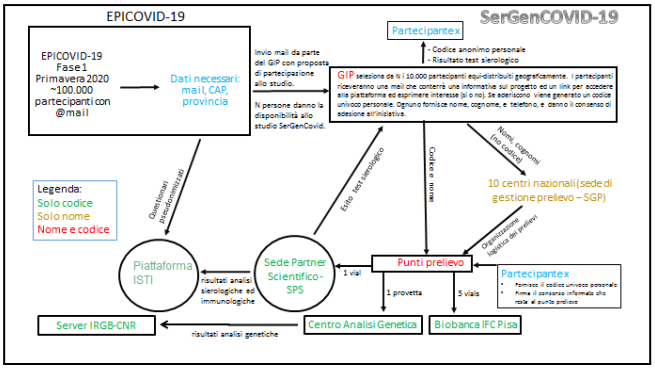


**Figura 1.** Schema a blocchi delle interazioni del progetto.

in questa fase per la realizzazione sia del backend che delle piattaforme *partecipante* e *operatore*. Come già accennato, all'interno del progetto l'ISTI è responsabile della raccolta, conservazione e consultazione dei dati sanitari (WP6), lavoro che viene svolto in stretta collaborazione con l'altro partner tecnologico del progetto lo IIT, il quale è responsabile della progettazione e nello sviluppo di una piattaforma informatica (GIP-Gestore Informatico della Piattaforma) che, oltre ad una descrizione generale del progetto, conterrà i dati dei soggetti coinvolti nella sperimentazione (vedi Figura 1).

### 1.1 Backend

Requisiti per realizzare l'applicazione web per l'accesso dei partecipanti:

- 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.

### 1.2 Dati

Uno dei requisiti fondamentali è che la piattaforma ISTI deve contenere solamente i dati anamnestici 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 e usato dalla piattaforma ISTI per gestire i dati memorizzati.

### 1.3 Interfaccia Partecipante

Il partecipante alla campagna 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.

| **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 |

**Tabella 1.** Tabella partecipante

| **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 |

**Tabella 2.** Tabelle questionario

**Accessi.** Per il partecipante l'accesso all'interfaccia ISTI è realizzata tramite il portale realizzato dallo IIT [1] usando le credenziali gestite da IIT.

**I dati.** I dati gestiti dalla piattaforma partecipante sono memorizzati nelle tabelle *partecipante* e *sgc19_quest* (vedi Tabella 1 e 2) della base di dati e sono:

- il codice univoco assegnato dalla piattaforma del WP5 a coloro che avranno aderito all'iniziativa;
- password per accedere ai referti dei test sierologici;
- i dati raccolti con il questionario anamnestico opportunamente cifrati;
- la data di acquisizione del questionario.

### 1.4 Interfaccia Operatore

L'operatore, cioè i responsabili per l'inserimento dei risultati dei test sierologici potranno:

- inserire per ogni partecipante, tramite il CUA, i referti dei test sierologici;
- consultare i sierologici inseriti per partecipante.

**Accessi.** Per l'operatore l'accesso all'interfaccia ISTI è realizzata tramite il portale D4Science [2].

**I dati.** I dati dei test sierologici per ogni partecipante, sempre senza alcun dato identificativo esplicito, ma identificati solo dal CUA, vengono memorizzati cifrati nella tabella *sierologico_test* della base di dati (vedi Tabella 3).
Quindi i dati contenuti e gestiti nella piattaforma *operatore* sono:

[1] https://sergencovid.iit.cnr.it
[2] https://www.d4science.org/

| **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 |

**Tabella 3.** Tabella risultati test sierologico.

- il CUA che identifica il partecipante;
- i dati specifici del referto del sierologico;
- la data di acquisizione dei risultati del sierologico.

## 2. Backend

In Figura 2 è mostrata l'architettura server side della piattaforma così come elaborata in fase di progettazione, con i seguenti elementi:

- load balancer ridondato 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, e accessibili solo dalle vm delle applicazioni web.

Per la parte di backend dell'applicazione web, si è optato di usare la metodologia, comunenemente usata, di Load Balancing che permette un'ottimizzazione dei carichi di lavoro per i server web in modo da offrire un servizio performante (con un servizio high performance) e affidabile nonchè reattivo ai guasti. Un Load Balancer può esso stesso essere ospitato da un server e permette di equilibrare il carico di lavoro tra i server garantendo meno rallentamenti e minori down di servizio assumendo il ruolo di controllore del traffico (Application Delivery Controller o ADC).

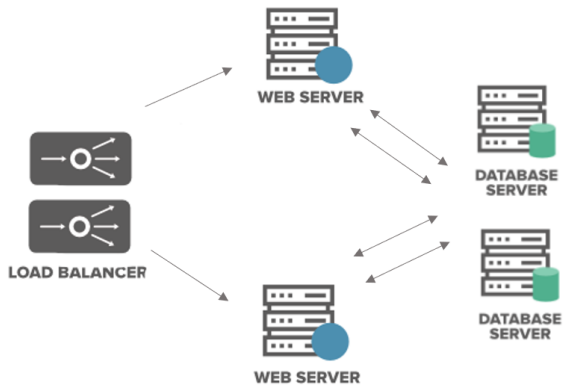


**Figura 2.** Backend

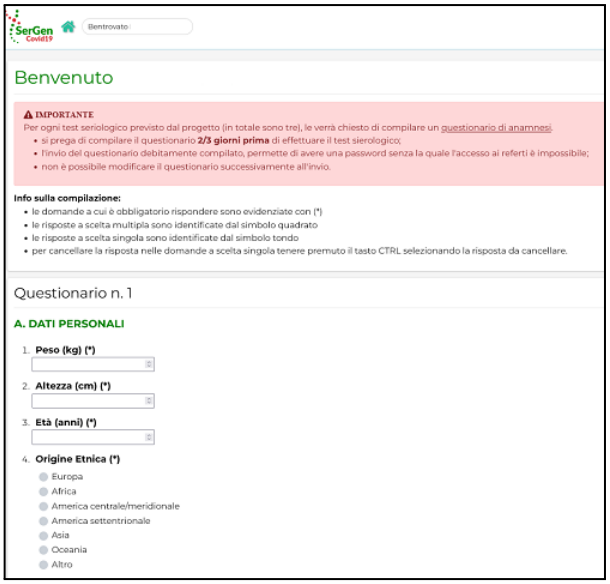


**Figura 3.** Interfaccia Partecipante: questionario da compilare.

Il software utilizzato per le funzionalità di load balancer è HAProxy[3], una soluzione open source veloce e affidabile. Ogni istanza è ridondata, e la High Availability tra le istanze HAProxy é gestita tramite keepalived[4].

Come server web si è optato per una soluzione open source tra le più conosciute: Nginx. Garantisce alte prestazioni ed efficienza degli applicativi. Popolare per la sua scalabilità, sicurezza e alta efficienza nell'uso delle risorse.

Per quanto riguarda invece la parte di Database Server, cosi come già deciso in fase di progettazione è stato usato PostgreSQL (versione 13) con l'estensione pg_crypto. PostgreSQL ha la capacità di crittografare e decrittografare elementi e impedendo ad altri di accedere a un valore di testo in chiaro: l'estensione pgcrypto in PostgreSQL fornisce funzioni e capacità crittografiche all'interno del database. Offre varie funzioni per eseguire operazioni crittografiche, tra cui crittografia, decrittografia, hashing e firme digitali. Nelle Tabelle 1, 2, 3 i dati criptati sono memorizzati come $Tipo = BYTEA$.

## 3. Realizzazione interfaccia partecipante

### 3.1 Interfaccia Web

Le interfacce web della piattaforma sono applicazioni web scritte in PHP con l'ausilio del template engine Smarty [5], il quale consente di separare l'interfaccia grafica di una pagina web dal back-end in PHP, favorendo lo sviluppo agile di applicazioni web e implementando il modello di sviluppo Model-View-Controller (MVC [6]), realizzando la componente (V)view dell'applicazione. In Figura 3 e 4 il risultato visuale.

[3] https://www.haproxy.org
[4] https://keepalived.org
[5] https://www.smarty.net
[6] https://martinfowler.com/eaaCatalog/modelViewController.html

### 3.2 Accesso

Ogni partecipante accede alla piattaforma ISTI per compilare il questionario anamnestico loggandosi sulla piattaforma GIP realizzata dallo IIT.
La suddetta transazione tra le due piattaforme, necessaria per mantenere separati i dati sensibili dai dati clinici (la prima piattaforma contiene solo i dati sensibili, e la seconda contiene solo i dati clinici), viene realizzata usando ua realtà ben consolidata conosciuta come JSON Web Token (JWT). Per lo scambio di informazioni tra i servizi viene generato un token che cifrato e firmato tramite una chiave disponibile solo a colui che lo ha effettivamente generato: il client invia una richiesta al server e questo genera un token di autenticazione che il client utilizzerà tutte le volte che andrà a collegarsi allo stesso nodo. Il JWT si compone dei seguenti campi:

- **Header**. Contiene il tipo di token e l'algoritmo di firma e/o crittografia utilizzato.
- **Payload**. Contiene le informazioni effettive che devono essere inviate all'applicazione il blocco che contiene le informazioni di scambio tra le parti. Questo a sua volta si divide in tre fasi: parametri registrati, parametri pubblici e parametri privati.
- **Signature**. La firma del JSON Web Token è creata utilizzando la codifica Base64 di Header e del Payload e il metodo di firma/codifica specificato. Alla fine dell'operazione viene generata una chiave che darà luogo a un token di oltre 200 caratteri.

Per l'implementazione dell'interfaccia partecipante del ISTI, il workflow realizzato è il seguente:

- accesso del partecipante sulla piattaforma GIP dello IIT tramite email e password;
- link con token verso la piattaforma ISTI, il token codifica il CUA del partecipante;
- sulla piattaforma ISTI decodifica del token per estrarre il CUA del partecipante;
- accesso all'area per la compilazione del questionario da parte del partecipante.

Per la corretta interpretazione del JWT, è stata utilizzata la libreria PHP-JWT[7].

[7] https://github.com/firebase/php-jwt

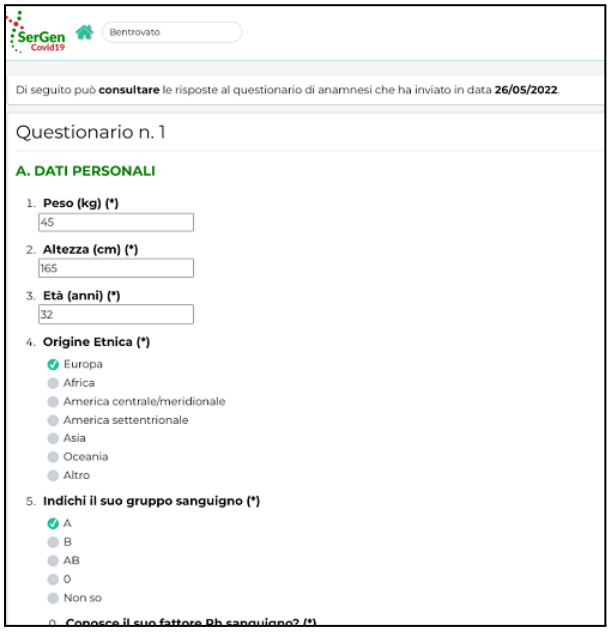


**Figura 4.** Interfaccia Partecipante: consultazione del questionario precedentemente inserito.

### 3.3 Questionario

Il questionario anamnestico, ideato dagli studiosi caratterizzare al meglio, sia anagraficamente che da un punto di vista anamnestico, i partecipanti al campione della popolazione generale partecipante al progetto.

La realizzazione della visualizzazione e del salvataggio del questionario ha comportato i seguenti passaggi preparatori:

- codifica per memorizzare in maniera compatta le domande e le risposte nella base di dati in modo da poter generare un file csv leggibile;
- codifica del questionario utile sia per trasformarlo in un formato html per renderlo fruibile sull'applicazione web che per trasformarlo in formato json per salvare il questionario nella base di dati.

**Codifica compatta.** Per poter memorizzare nella base di dati l'intero questionario in un formato compatto si è realizzato un mapping tra le domande testuali e una codifica composta da lettere e numeri, e le risposte sono state mappate a loro volta come un valore numerico.

In Figura 5 è mostrato un esempio di questa codifica: per la domanda *È stato vaccinato contro il Coronavirus (SARS-CoV-2)?* la codifica della domanda è il codice **B2**, e per le sottodomande relative *In quale categoria rientra?*, *È a conoscenza del tipo di vaccino che Le è stato somministrato?* si hanno rispettivamente i codici **B21** e **B22**.

```
[...]
B2 È stato vaccinato contro il Coronavirus
 ↪  (SARS-CoV-2)?
 - No
 - Si
        B21 In quale categoria rientra?
                o Degente Residenza sanitaria
                 ↪  assistenziale (RSA), Hospice, Case
                 ↪  di riposo
                o Docente
                o Età oltre 80 anni
                o Farmacista / Biologo
                o Operatore sanitario (medico,
                 ↪  infermiere, Operatore Socio
                 ↪  Sanitario)
                o Persone a rischio per malattie
                o B21T Altro_____________
        B22 È a conoscenza del tipo di vaccino che Le è
         ↪  stato somministrato?
                o Moderna
                o Oxford - AstraZeneca
                o Pfizer-BioNTech
                o Non ricordo
                o B22T Altro vaccino_____________
[...]
```

**Figura 5.** Esempio di codifica compatta per le domande questionario.

In Figura 6 è mostrato un esempio della codifica delle risposte: per la domanda con codice **B2** le possibile risposte {*No, Si*} vengono codificate numericamente come {**1,2**}, per la domanda **B21** le possibile risposte {*Degente Residenza sanitaria, Docente, ... ,Persone a rischio per malattie, Altro*} vengono codificate numericamente come {**1,2,...,6,7**}.

**Codifica di visualizzazione.** Per poter visualizzare in maniera efficiente il questionario fornito dagli esperti in formato html, l'intero questionario è stato mappato in un file XML e tramite un parser PHP realizzato ad hoc (*XMLSurvey.php*) viene generato il codice HTML. In Codice 1 è mostrato un esempio di questa codifica: per la domanda alla riga 6 abbiamo risposte a scelta esclusiva indicate dall'elemento <*answers type="single*"> e il parser le trasforma in Radio button in HTML, mentre alla domanda di riga 27 abbiamo delle risposte a scelta multipla indicate dall'elemento <*answers type="multiple*"> e il parser le trasforma come checkbox in HTML.

Al momento della sottomissione del questionario, viene generato un array in POST con le associazioni $k, v$ di domanda, risposta (così come esposto nel paragrafo successivo), poi convertito in formato JSON memorizzato nella base di dati.

**Codice 1.** Esempio codifica XML del questionario.

```
<question label="È stato vaccinato contro il Coronavirus (SARS-CoV-2)?">
 <answers type="single">
  <answer label="No" />
  <answer label="Si">
   <questions>
    <question label="In quale categoria rientra?">
     <answers type="single">
      <answer label="Degente Residenza sanitaria assistenziale (RSA), Hospice, Case di riposo" />
      <answer label="Docente" />
      <answer label="Età oltre 80 anni" />
      <answer label="Farmacista / Biologo" />
      <answer label="Operatore sanitario (medico, infermiere, Operatore Socio Sanitario)" />
      <answer label="Persone a rischio per malattie " />
      <answer label="Altro" customAnswer="true" />
     </answers>
    </question>

    <question label="È a conoscenza del tipo di vaccino che Le è stato somministrato?">
     <answers type="single">
      <answer label="Moderna" />
      <answer label="Oxford - AstraZeneca" />
      <answer label="Pfizer-BioNTech" />
      <answer label="Non ricordo" />
      <answer label="Altro vaccino" customAnswer="true"/>
     </answers>
    </question>
 <question label="A quali trattamenti è stato sottoposto per la cura del COVID-19?">
  <answers type="multiple">
   <answer label="Antipiretici o antidolorifici" />
   <answer label="Azitromicina" />
   <answer label="Clorochina" />
   <answer label="Cortisonici" />
   <answer label="Eparina a basso peso molecolare" />
   <answer label="Lopinavir/ritonavir (Kaletra)" />
   <answer label="Ossigenoterapia" />
   <answer label="Remdesevir" />
   <answer label="Altri antibiotici" />
   <answer label="Non lo so" />
   <answer label="Nessuno" clear="true"/>
   <answer label="Altro" customAnswer="true"/>
  </answers>
 </question>
```

```
[...]

B2 È stato vaccinato contro il Coronavirus
 ↪  (SARS-CoV-2)?
 1 No
 2 Si
        B21 In quale categoria rientra?
                1 Degente Residenza sanitaria
                 ↪  assistenziale (RSA), Hospice, Case
                 ↪  di riposo
                2 Docente
                3 Età oltre 80 anni
                4 Farmacista / Biologo
                5 Operatore sanitario (medico,
                 ↪  infermiere, Operatore Socio
                 ↪  Sanitario)
                6 Persone a rischio per malattie
                7 Altro__________
[...]
```

**Figura 6.** Esempio di codifica compatta per le risposte questionario.

Il questionario compilato viene registrato nella colonna *risposte* della tabella *sgc19_quest* (vedi Tabella 2) come:

$$questionario = \{(k,v) : \begin{array}{l} \\ k \text{ codifica domanda,} \\ v \text{ codifica risposta e HTML}\} \end{array}$$

cioè un insieme di coppie (chiave, valore) dove la chiave è la codifica della domanda e il valore è la codifica delle risposte: in Codice 2, alla chiave ["*B22*"] viene associata il valore 2 di ["*answer_index*"] ( ["*question*"] e ["*answer*"] sono i valori per la generazione testuale nei campi HTML).

**Codice 2.** Esempio di codifica JSON del questionario.

```
["B22"]=>
{
  ["question"]=> "È a conoscenza del tipo di vaccino che
      Le è stato somministrato?"
  ["answer"]=> "Oxford - AstraZeneca"
  ["answer_index"]=> 2
}
```

## 4. Realizzazione interfaccia operatore

A differenza dell'interfaccia partecipante che è strettamente correlata con la piattaforma GIP dello IIT, l'interfaccia operatore si appoggia sull'infrastruttura D4Science che fornisce autenticazione e autorizzazione e VRE[8] per poter realizzare la piattaforma.

[8] https://datascience.codata.org/articles/10.2481/dsj.GRDI-013

### 4.1 Interfaccia Web

Le interfacce web della piattaforma sono scritte in PHP 7.4 con l'ausilio del template engine Smarty e sono riversate in una portlet D4Science. Le portlet sono un tipo speciale di servlet, moduli web progettati per essere inseriti facilmente in un portale web.

### 4.2 Accesso

Questa interfaccia usufruisce del sistema di autenticazione in uso sull'infrastruttura D4Science, che permette di accedere al gateway D4science appositamente creato per il progetto [9], usando le credenziali che preferisce tra quelle supportate (Accademiche, Google, Twitter, LinkedIn, eccetera). Ogni operatore che si registra al gateway del progetto, riceve dall'amministratore abilitato le autorizzazioni ad operare sulla piattaforma. Autenticazione e autorizzazione sono a carico del gateway D4Science, e la piattaforma operatore riceve da questo un token JWT con le informazioni necessarie per dare accesso all'interfaccia web tramite la quale l'operatore potrà inserire i risultati dei test sierologici (vedi Figura 7 e 8).

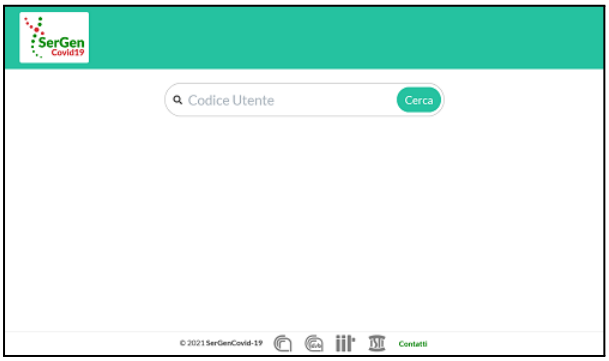


**Figura 7.** Interfaccia operatore: schermata principale.

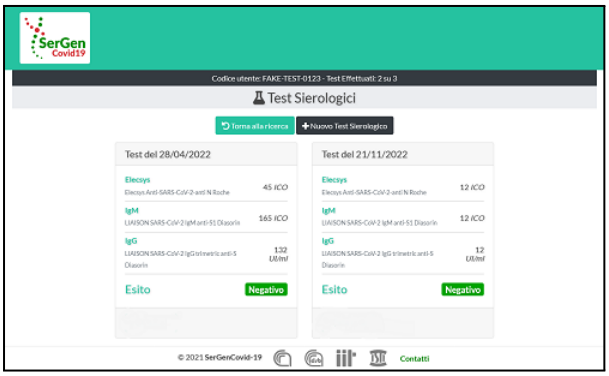


**Figura 8.** Interfaccia operatore: visualizzazione e inserimento test sierologico.

## 5. Conclusioni

Per la corretta e efficiente realizzazione delle due piattaforme sono stati creati:

- server di produzione: utilizzato per distribuire le piattaforme, sottoposte a sviluppo e test approfonditi prima di essere convalidati come pronti per la produzione;
- server di sviluppo: progettato per facilitare lo sviluppo e i test delle piattaforme;

sempre ridondati. Provisioning, configurazione, e deployment delle applicazioni sia sui server di produzione che quelli di sviluppo sono effettuati tramite Ansible[10], come tool di automazione. La piattaforma partecipante è stata rilasciata on-line il 28 Maggio 2021, mentre la piattaforma operatore il 6 Luglio 2021. In fase di running delle due piattaforme, non si sono evidenziati problemi, se non quelli dovuti a variazioni dei requisiti e a modifiche del questionario in corso d'opera: in entrambi i casi, data la ridondanza dell'architettura, non ci sono stati downtime delle piattaforme.

[9] https://sergencovid19.d4science.org/

[10] https://www.ansible.com/