UNIVERSITÀ DEGLI STUDI DI CAGLIARI

FACOLTÀ DI SCIENZE

Corso di Laurea Triennale in Informatica

Analisi ed estrazioni di informazioni da referti medici: Un caso di studio e implementazione nel progetto SIMIOR

Relatore
Dott. Gianmarco Cherchi

Studente
Lorenzo Ludovico Concas
Matr. N. 65315

ANNO ACCADEMICO 2021/2022

In the treatment path of hospital patients, one of the fundamental steps to define the state of health is the execution of tests aimed at better understanding the pathologies encountered. This information is expressed in medical reports, which together determine the progression of the patient in his hospital career. But keeping track of these trends can be complicated by human error in transcribing the information.

This thesis analyses the methodology implemented in the SIMIOR project to solve the problem and its future developments

Nel percorso di cura dei pazienti ospedalieri, uno dei passaggi fondamentali per definire lo stato di salute è l'esecuzione di test mirati a comprendere meglio le patologie riscontrate. Queste informazioni sono espresse in referti medici, il cui insieme determina la progressione del paziente nel suo percorso ospedaliero. Tenere traccia di questi andamenti può però essere complicato dall'errore umano nella trascrizione delle informazioni.

Questa tesi analizza la metodologia implementata nel progetto SIMIOR per risolvere il problema e i relativi futuri sviluppi

Indice

Capitolo 1

Introduzione

1.1 SPIN-UTI

L'acronimo SPIN-UTI (Simplified Prognostic INSevere Sepsis in ICU) indica un sistema di punteggio pensato per valutare il rischio di mortalità dei pazienti affetti da sepsi grave ricoverati nei reparti di terapia intensiva (UTI).
Questo sistema di punteggio, in uso nel reparto di terapia intensiva del Presidio Ospedaliero Duilio Casula è implementato a livello informatico con l'utilizzo dei software Microsoft Excel e Access, il primo utilizzato per inserire le informazioni e per effettuare i calcoli tramite formule, il secondo utilizzato come base dati. Viene logico intuire l'inidoneità del sistema attuale, soprattutto sul fronte organizzativo dei dati.

1.2 Il progetto SIMIOR

Il Simior è un sistema informatico creato a inizio 2022 per sopperire alle limitazioni dell'attuale implementazione del sistema di punteggio SPIN-UTI attualmente utilizzato nel già citato reparto di terapia intensiva. I destinatari di questo progetto sono i dottori, denominati utenti, che accedendo tramite pagina web potranno inserire le cartelle cliniche dei pazienti in cura, potendo tracciare l'andamento del ricovero e le statistiche del reparto.

1.3 Differenze con l'implementazione attuale

La principale limitazione dell'implementazione attuale di SPIN-UTI è il limite di informazioni inseribili per ogni scheda-paziente, con la conseguente duplicazione delle stesse al fine di memorizzare tutte le informazioni sulla la degenza.

Ne risulta un sistema poco ordinato e maggiormente soggetto a errori durante la copia delle informazioni essenziali.
Il simior è modellato sulle esigenze specifiche del reparto e viene adattato di conseguenza all'evolversi delle necessità. La necessità che ha portato alla stesura di questa tesi è l'estrazione delle informazioni dai referti di laboratori, nello specifico gli antibiogrammi. Questa funzionalità permette, in maniera semplice per l'operatore, di allegare un referto in formato PDF ed ottenere in una tabella, visivamente analoga ad altre parti del progetto, i dati contenuti nel documento originale.

1.4 Dettagli tecnici sul SIMIOR (Da togliere?)

Il SIMIOR è implementato come sito web scritto in java in esecuzione sul server open-source GlassFish, ospitato sui servizi AWS di Amazon. Il front-end è implementato con un mix di tecnologie quali JSP (per la gestione delle pagine e dei dati) e Bootstrap per l'impaginazione dei contenuti.

Capitolo 2

I referti medici

2.1 La struttura dei referti

I referti prodotti nei laboratori del policlinico seguono la struttura standard adottata dall'Azienda Tutela Salute Sardegna (dal 2022 Azienda Regionale della Salute, ARES) presentando una divisione in 4 blocchi: Se sono presenti più analisi, o le informazioni dell'analisi superano una certa quantità (per esempio un antibiogramma molto lungo) il documento viene suddiviso su più pagine. Di norma i referti prodotti non superano le due pagine.

- Intestazione ATS
- Sezione anagrafica
- Contenuto referto
- Piè di pagina

2.1.1 L'intestazione ATS

Il primo blocco, contenente l'intestazione, identifica la provenienza del documento mostrando informazioni sull'ASL quali strutture ad essa collegate, i recapiti telefonici e il logo. La presenza di queste informazioni è uno dei requisiti per confermare la validità del documento inserito nel SIMIOR.

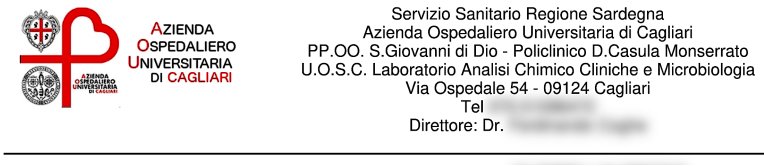


Figura 2.1: *Intestazione iniziale di un referto*

2.1.2 La sezione anagrafica

Segue la sezione anagrafica che indica la priorità dell'analisi effettuata, la provenienza del paziente (intesa come reparto di provenienza) e le informazioni personali del paziente.

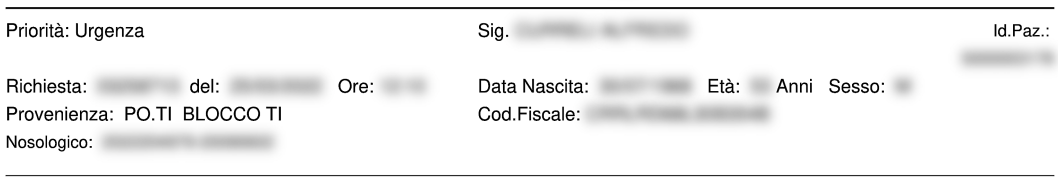
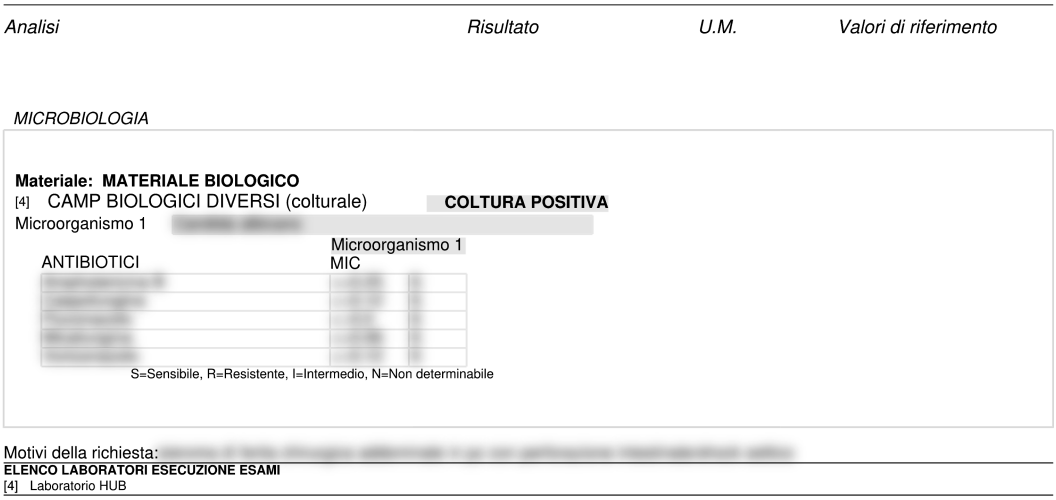


Figura 2.2: *Sezione anagrafica che mostra la priorità e le informazioni sul paziente*

La presenza di queste informazioni consente di verificare che il referto inserito nel sistema sia effettivamente del paziente selezionato, prevenendo cosi erronea associazione.

2.1.3 Il contenuto del referto

Nel cuore del referto è collocato il risultato dell'analisi, che varia a seconda dell'obbiettivo del test. L'implementazione attuale è progettata per trovare ed estrarre le tabelle antibiogramma.


Analisi *Risultato* *U.M.* *Valori di riferimento*

MICROBIOLOGIA

Materiale: MATERIALE BIOLOGICO
[4] CAMP BIOLOGICI DIVERSI (colturale) **COLTURA POSITIVA**
Microorganismo 1

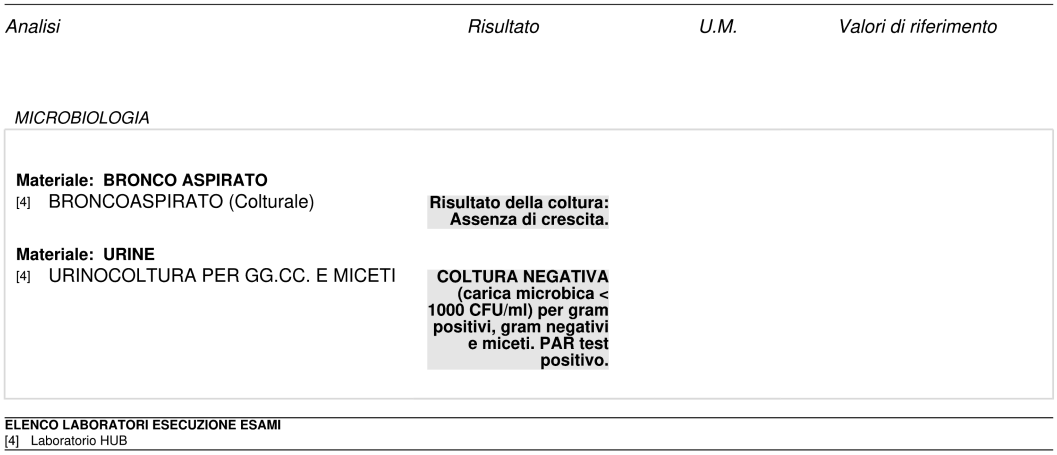
ANTIBIOTICI	Microorganismo 1 MIC	
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

S=Sensibile, R=Resistente, I=Intermedio, N=Non determinabile

Motivi della richiesta: [illegible]
ELENCO LABORATORI ESECUZIONE ESAMI
[4] Laboratorio HUB

Figura 2.3: *Un antibiogramma valido per l'estrazione*

Nei referti validi, è presente una o più tabelle antibiogramma costituita da un minimo di tre colonne e un massimo indefinito ma solitamente non più di sette (una per gli antibiotici e due per ogni microrganismo fino a tre). La prima colonna, comune a tutti i microrganismi a cui la tabella fa riferimento, indicano l'antibiotico testato e valori risultanti. Segue la colonna la cui intestazione riporta la parola "MIC", che indica la "Minima Concentrazione" ossia la concentrazione più bassa (solitamente espressa come μg/mL) per cui un determinato antibiotico è capace di inibire la crescita del microrganismo associato. Dopo il MIC segue la colonna che rappresenta la sensibilità del microrganismo all'antibiotico testato con un set di valori definiti nella legenda posta sotto la tabella. L'unione del microrganismo, gli antibiotici e relativi mic e sensibilità forma l'antibiogramma per il microrganismo in questione.



Analisi	*Risultato*	*U.M.*	*Valori di riferimento*
MICROBIOLOGIA			
Materiale: BRONCO ASPIRATO			
[4] BRONCOASPIRATO (Colturale)	**Risultato della coltura: Assenza di crescita.**		
Materiale: URINE			
[4] URINOCOLTURA PER GG.CC. E MICETI	**COLTURA NEGATIVA (carica microbica < 1000 CFU/ml) per gram positivi, gram negativi e miceti. PAR test positivo.**		

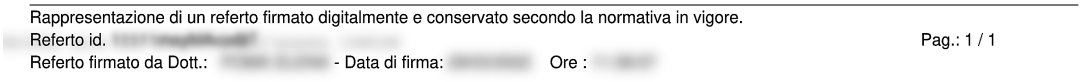
ELENCO LABORATORI ESECUZIONE ESAMI
[4] Laboratorio HUB

Figura 2.4: *Un referto non valido per l'estrazione delle tabelle*

Nella figura 2.4 viene mostrata la sezione contenuto di un referto dove non sono presenti tabelle estraibili, pertanto non valido ai fini delle funzionalità implementata.

2.1.4 Il piè di pagina

Infine, il piè di pagina, che contiene informazioni sulla firma del documento, la relativa data e l'ora. Inoltre è presente un codice che identifica il referto negli altri sistemi del policlinico.



Rappresentazione di un referto firmato digitalmente e conservato secondo la normativa in vigore.
Referto id. [redacted] Pag.: 1 / 1
Referto firmato da Dott.: [redacted] - Data di firma: [redacted] Ore : [redacted]

Figura 2.5: *Piè di pagina (Footer)*

2.2 Tipologie di referti

Nell'ambito di questa tesi possiamo distinguere 5 tipologie di referti a seconda del loro contenuto:

- Antibiogramma con singolo microrganismo
- Antibiogramma con più microrganismi
- Antibiogramma singolo multi-pagina
- Antibiogramma multiplo multi-pagina
- Referto senza tabelle

La prima tipologia è costituita da un singolo microrganismo seguito da una tabella con sole tre colonne, strutturate come descritto nel paragrafo 2.1.3, la figura 2.3 mostra nel dettaglio una tabella ad antibiogramma singolo.

L'antibiogramma multiplo (con più microrganismi) invece presenta più microrganismo accorpati nella stessa tabella, pertanto avremmo una sola colonna per l'antibiotico ma un susseguirsi di colonne MIC-Sensibilità per ciascun microrganismo. Le tabelle multipagina costituiscono una variazione rara ma non impossibile delle prime due tipologie, presentando le tabelle fisicamente divise fra una o più pagine.

2.2.1 La sorgente dei dati

I PDF dei referti sono generati tramite una libreria software chiamata "iText" che riceve in ingresso i dati degli antibiogrammi, sorge dunque spontanea la domanda:

"Perchè non aggirare il referto e prelevare i dati dalla sorgente?"

Purtroppo i sistemi informatici del policlinico sono isolati e non di facile modifica, soprattutto a livello legale. Un eventuale alterazione del prodotto software designato implicherebbe un eventuale perdita di supporto dalla casa madre e problemi di privacy. Oltretutto potrebbe essere necessario inserire referti non generati dallo specifico programma utilizzato in un determinato reparto.

Capitolo 3

Il PDF e l'Estrazione delle informazioni

Prima di poter estrarre dati utili dai documenti PDF è necessario operare delle trasformazioni o delle rappresentazioni, dato che, per natura del PDF non vi sono dei dati utili al nostro scopo direttamente leggibili.

3.1 Cosa sono i PDF?

PDF è un acronimo che sta per *Portable Document Format* è uno formato standard creato nel 1993 allo scopo di rappresentare documenti testuali o con immagini indipendentemente dall'hardware o dal software che usato per generare o visualizzare il documento stesso. I file costruiti in questo formato sono rappresentati tramite una sequenza di caratteri ASCII, all'interno possiamo individuare quattro strutture:

- Header
- Body
- Cross-Reference Table
- Trailer

Ciò che viene presentato visivamente all'utente è contenuto nella sezione `body` del PDF, su questo flusso di caratteri si opera per ottenere le informazioni volute.

3.2 Tecniche di estrazione

Le tecniche di estrazione sono principalmente due:

- Conversione in formati testuali
- Lettura diretta del flusso `Body`

Nel primo caso si cerca di convertire (tramite librerie o servizi appositi) il documento in un file di testo dove poter effettuare l'estrazione. Questo metodo è stato scartato durante i test d'implementazione preliminari per via del risultato insoddisfacente, difatti i tentativi di conversione (con strumenti già esistenti e librerie dedicate) non riuscivano a convertire le tabelle in testo ma venivano generate delle immagini. Non è stato possibile utilizzare librerie dedicate all'estrazione di tabelle poiché gli antibiogrammi hanno una posizione fissa. Nel secondo caso la realizzazione di un prodotto completo avrebbe richiesto studi molto più lunghi e fuori dallo scopo della tesi.

3.3 La libreria utilizzata

In un primo momento si è tentato di utilizzare la stessa libreria responsabile della generazione dei documenti, ma limitazioni di licenza e funzionalità hanno favorito la concorrente open-source Apache PDFBox che si è rivelata oltretutto più versatile e adatta allo scopo prefissato.

3.4 L'approccio iniziale e le problematiche

Il primo approccio si è basato sulla semplice estrazione del testo linea per linea dai referti con l'applicazione di specifiche regex. Avendo inizialmente soltanto un referto su cui effettuare i test non è apparso subito l'evidente problema sorto in seguito con gli antibiogrammi multi-microrganismo. L'algoritmo di estrazione era dunque costituito dai seguenti passaggi:

1. Scorrere la lista fino a trovare la stringa corrispondente alla parola "Microrganismo 1" seguito dal nome del microrganismo
2. Scorrere di 2 posizioni (equivalenti a saltare la riga "Microrganismo 1" e "ANTIBIOTICO MIC")

3. Verificare ed estrarre le informazioni riguardo alla riga della tabelle tramite la seguente regex:

```
/(?<NOME>[A-z.]+)(?<MIC><*=*[0-9]*[,0-9]*)*(?<SENSIBILITA>IE{1}|[(SRIN]{1})/g
```

Figura 3.1: *Regex per l'estrazione delle righe della tabella*

4. Ripetere fino all'arrivo della riga contenente la legenda.

Per poter estrarre le informazioni da tabelle con più microrganismi è possibile modificare la regex ripetendo la seconda riga (la parte col gruppo MIC e Sensibilità) quante volte sono i microrganismi individuati. Questa soluzione però fallisce nel caso di coppia di celle vuote (che significa analisi antibiotica non condotta per tale microrganismo).

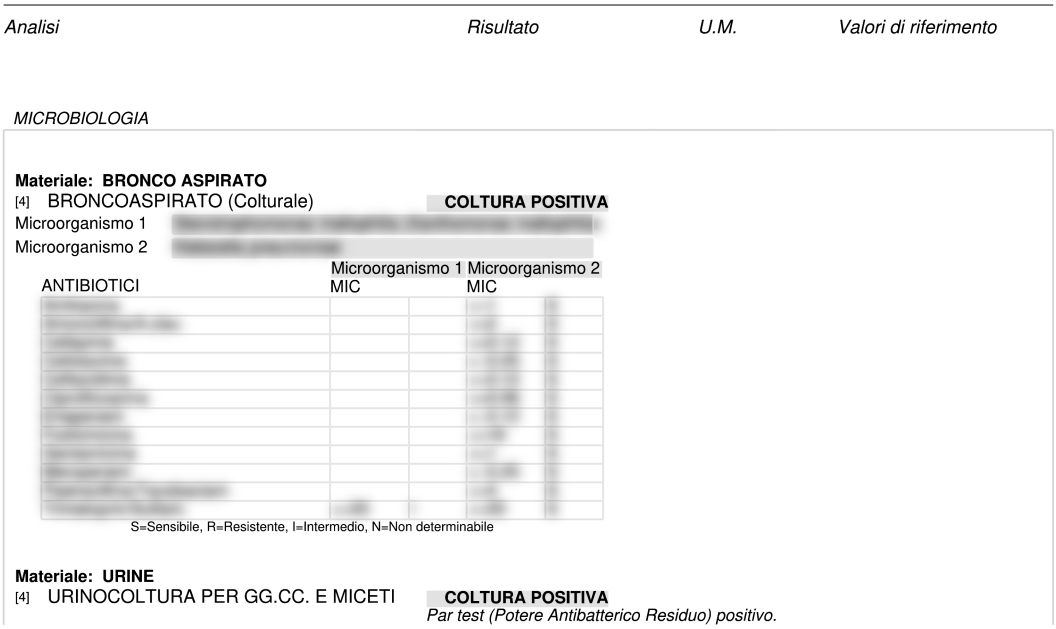


Figura 3.2: *Tabella con più microrganismi*

Per esempio, nella tabella mostrata nella figura 3.3 soltanto l'ultima riga potrà estrarre dei risultati validi tramite modifica delle regex, che risulterebbe la seguente: Dove *NR_MICRO* è da sostituire col numero di microrganismi.

```
/(?<NOME>[A-z.]+){NR_MICRO}((?<MIC><*=*[0-9]*[,0-9]*)*(?<SENSIBILITA>IE{1}|[(SRIN]{1}))/g
```

Figura 3.3: *Regex per l'estrazione delle righe della tabella*

3.5 Il secondo approccio

Il secondo approccio, che apporta la correzione più significativa rivede completamente la logica di estrazione e aggiunge delle modifiche alla libreria PDFBox. Il concetto fondamentale di quest'approccio è un estrazione "ibrida" ossia si ricerca l'inizio della tabella come nel metodo precedente, ma l'estrazione dei dati vera è propria viene effettuata calcolando la posizione delle varie celle.
Infatti anche se la posizione e la dimensione della tabella non è costante, le celle invece lo sono, permettendo un calcolo preciso della loro posizione e conseguente estrazione.

3.5.1 Modifica della libreria

Per ottenere la posizione degli elementi si è proceduto ad estendere la classe *PDFTextStripper* della libreria aggiungendo prima una lista di oggetti di tipo `<string, List<TextPosition>`, ogni oggetto di questa lista è composto da una coppia chiave-valore in cui la chiave è una qualsiasi parola estratta e la chiave è una lista di posizioni in cui ogni lettera ha una sua coordinata. Come seconda cosa è stato modificato il metodo *writeString* in modo che inserisse nella lista prima citata queste nuove informazioni lasciando però inalterato il resto delle istruzioni.

3.5.2 Localizzazione delle celle

Una volta generate queste informazioni è possibile rappresentare l'algoritmo come segue:

1. Si effettua una prima ricerca della tabella in modo analogo al metodo precedente
2. Si continua a scorrere la lista e si tiene traccia del numero di microrganismi presenti nella tabella trovata
3. Finita l'enumerazione dei microrganismi, si inizia a scorrere la nuova lista della classe modifica fino a trovare la parola chiave "ANTIBIOTICI", seguita da tante stringhe "MIC" quanti sono i microrganismi
4. L'elemento puntato dalla lista sarà il nome dell'antibiotico, da qui si calcolano le dimensioni e le posizioni delle celle MIC e Sensibilità
5. Definiamo tre "regioni" (dei rettangoli in cui operare), una per estrarre il nome dell'antibiotico alla linea selezionata, una per quella successiva e una per la legenda

6. Si tenta la prima estrazione del testo, seguita poi, per ogni microrganismo dai seguenti passi:

   (a) Si calcola la cella MIC le cui coordinate saranno: distanza colonna antibiotico $+156+$ (n° microrganismo)$*42$

   (b) Discorso analogo per la cella Sensibilità che ha coordinate : distanza colonna MIC + (n° microrganismo) $*30*$

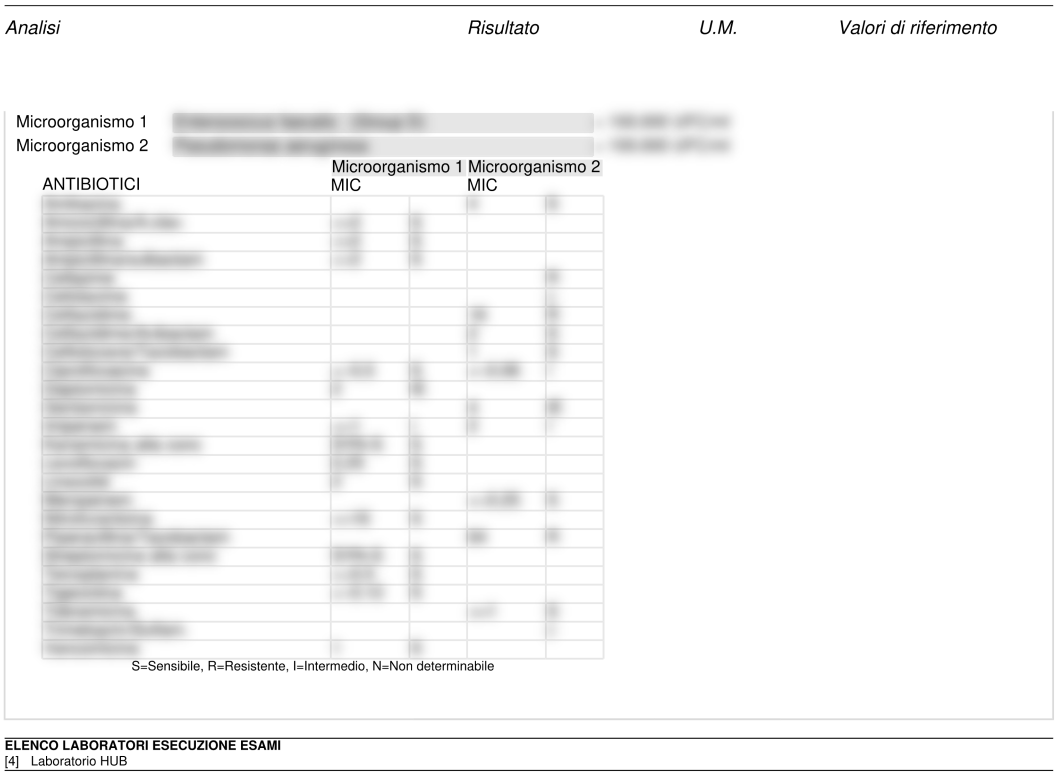
   (c) Vengono definite altre due regioni, si tenta l'estrazione e si inseriscono le informazioni raccolte nell'antibiogramma

7. Fatto questo, si scorre la lista delle parole fino a trovare o la legenda (che indica la fine della tabella e dell'estrazione) o il prossimo antibiotico (si confronta il testo con il dato estratto prima). Nel secondo caso si ferma lo scorrere della lista e si riparte del punto 4.

Questi 7 passaggi sono ripetuti per ogni pagina del PDF, ed è una soluzione quasi definitiva ma non esente da difetti. Ma cosa succede se ci sono più tabelle per pagina? Semplicemente viene rilevata soltanto la prima tabella e le successive vengono ignorate.

3.6 La soluzione finale

Al fine di ridurre la complessità del codice, è stato effettuata una riorganizzazione e parziale riscrittura che ha portato alla suddivisione del codice in due sotto-funzioni principali: la rilevazione delle tabelle e l'estrazione. La rilevazione delle tabelle effettua un controllo pagina per pagina e differentemente da prima la ricerca continua fino all'indice delle line che rappresenta il fine pagina. L'estrazione delle tabelle funziona in modo analogo alla soluzione precedente, con l'unica differenza che viene invocata con gli indici di inizio e di fine della tabella su cui operare. Questi piccoli accorgimenti permettono una migliore lettura del codice e di estrarre tutte le tabelle presenti, non soltanto la prima, rendendo l'algoritmo più preciso. Non sono ancora stati testati referti con tabelle divise su più pagina perché non sono stati forniti esempi, si pensa che esistano perché nei campioni forniti alcuni presentano informazioni su materiale d'estrazione o componenti grafici in pagine differenti rispetto alla posizione della relativa tabella. Un esempio è la figura 3.3 dove avviene proprio l'esempio del materiale, infatti la tabella relativa si trova nella seconda pagina



Analisi	Risultato	U.M.	Valori di riferimento

Microorganismo 1 [redacted]
Microorganismo 2 [redacted]

ANTIBIOTICI	Microorganismo 1 MIC		Microorganismo 2 MIC	
[redacted]				

S=Sensibile, R=Resistente, I=Intermedio, N=Non determinabile

ELENCO LABORATORI ESECUZIONE ESAMI
[4] Laboratorio HUB

Figura 3.4:

La tabella in questione non presenta appunto il materiale su cui è basata l'analisi, poichè non è noto l'algoritmo utilizzato per dividere le informazioni si propone una soluzione aggiuntiva, da verificare una volta ottenuti campioni conformi, nella seguente forma algoritmica:

1. Per ogni pagina, rimuovere le sezioni:
   (a) *Intestazione* (fig. 2.2)
   (b) *Anagrafica* (fig. 2.3)
   (c) *Piè di Pagina* (fig. 2.3)
2. Unire le pagina in una sola
3. Procedere dal punto 1) dell'algoritmo proposto nel paragrafo 3.5.2

Capitolo 4

Le modifiche al SIMIOR

Per accomodare le nuove funzionalità è stato necessario fare delle modifiche strutturali al SIMIOR, spaziando dal database a classi di gestione dei dati interne al front-end del progetto.

4.1 Modifiche strutturali

Le modifiche fatte si possono raggruppare in tre punti:

- Database
- Archivio Statico
- Front-End

Oltre al database che contiene i dati creati dagli utenti, il SIMIOR possiede dei dati "statici" che raramente richiedono una modifica, un esempio di questi dati sono i codici degli antibiotici definiti dal sistema ATC/DDD (Anatomical Therapeutic Chemical/Defined Daily Dose) o i codici dei microorganismi definiti dall'Istituto Superiore di Sanità. Quando viene effettuato un inserimento di un antibiogramma (sia tramite la nuova funzionalità sia manualmente) il sistema effettua una ricerca in questo archivio statico per estrarre il codice relativo al microrganismo/antibiotico, ma inizialmente questi nomi erano in lingua inglese (sempre seguendo il sistema di origine dei dati) causando il fallimento dell'estrazione. Per ovviare a questo problema si è prima provveduto ad aggiornare l'archivio con la versione italiana e in secondo luogo aggiungendo la possibilità di avere un valore alternativo per casi particolari verificati in alcuni referti. Che altro ho cambiato?

4.2 Implementazione FrontEnd

L'utente può utilizzare la funzionalità recandosi nella sezione *infezione, contaminazione o colonizzazione* di un qualsiasi ricovero, dove troverà una scheda contenente l'essenziale per poter allegare un referto e procedere con il caricamento.

4.2.1 Localizzazione della funzionalità

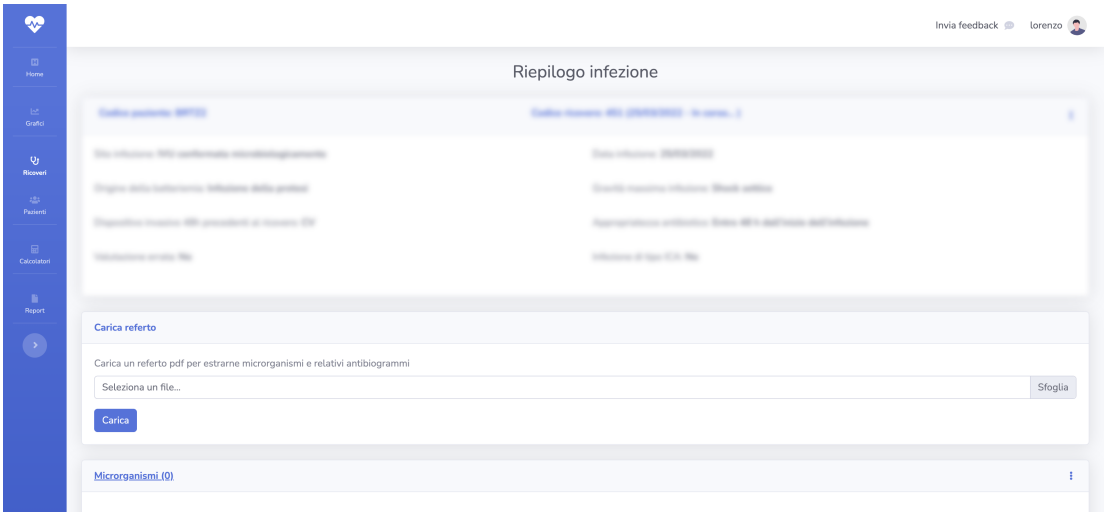


Figura 4.1: *Scheda upload documento*

La pressione del tasto *"Sfoglia"* aprirà una schermata di dialogo gestita dal sistema che permetterà di selezionare il file determinato. Fatto questo l'utente procede alla pressione del tasto *Carica* che lo porterà a una seconda pagina dove verranno mostrati i risultati dell'estrazione.

4.2.2 La schermata dei risultati

Capitolo 5

Conclusioni

5.1 Risultati

5.2 Sviluppi futuri