



**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



A scoring system is in place in Italy's intensive care units (ICUs) used to assess the risk of mortality in patients with severe sepsis. There is currently no standardized computer system for calculating this score, consequently the Duilio Casula Hospital Hospital in collaboration with the Faculty of Sciences of the University of Cagliari have collaborated to create the SIMIOR project. To calculate this score and extrapolate statistics, however, data from medical records is required, which is entered manually.

This thesis analyzes the implementation of the necessary automatisms to accelerate the process of entering the medical reports into the system, the structure of the medical reports and the changes implemented to the SIMIOR system.



Nei reparti di terapia intensiva italiani (UTI) è in vigore un sistema di punteggio utilizzato per valutare il rischio di mortalità nei pazienti affetti da sepsi grave. Non esiste attualmente un sistema informatico standardizzato per il calcolo di questo punteggio, di conseguenza il Policlinico Ospedaliero Duilio Casula in collaborazione con la Facoltà di Scienze dell’Università di Cagliari hanno collaborato per creare il progetto SIMIOR. Per calcolare questo punteggio ed estrapolare statistiche sono però necessari dati delle cartelle cliniche, che vengono inseriti manualmente.

Questa tesi analizza l’implementazione dei necessari automatismi per accelerare il processo di inserimento dei referti medici nel sistema, la struttura dei referti medici e le modifiche attuate al sistema SIMIOR.



# Indice

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduzione</b>                                | <b>1</b>  |
| 1.1      | SPIN-UTI . . . . .                                 | 1         |
| 1.2      | Soggetto di studio . . . . .                       | 2         |
| <b>2</b> | <b>SPIN-UTI e SIMIOR</b>                           | <b>3</b>  |
| 2.1      | Implementazione di SPIN-UTI . . . . .              | 3         |
| 2.2      | Limitazioni del sistema attuale . . . . .          | 3         |
| 2.3      | Il progetto SIMIOR . . . . .                       | 4         |
| 2.4      | Differenze con l'implementazione attuale . . . . . | 4         |
| <b>3</b> | <b>I referti medici</b>                            | <b>5</b>  |
| 3.1      | La struttura dei referti . . . . .                 | 5         |
| 3.1.1    | L'intestazione ATS . . . . .                       | 6         |
| 3.1.2    | La sezione anagrafica . . . . .                    | 6         |
| 3.1.3    | Il contenuto del referto . . . . .                 | 7         |
| 3.1.4    | Il piè di pagina . . . . .                         | 8         |
| 3.2      | Tipologie di referti . . . . .                     | 9         |
| 3.3      | La sorgente dei dati . . . . .                     | 10        |
| <b>4</b> | <b>Il PDF e l'Estrazione delle informazioni</b>    | <b>11</b> |
| 4.1      | Cosa sono i PDF? . . . . .                         | 11        |
| 4.2      | Tecniche di estrazione . . . . .                   | 12        |
| 4.3      | La libreria utilizzata . . . . .                   | 12        |
| 4.4      | L'approccio iniziale e le problematiche . . . . .  | 12        |
| 4.5      | Il secondo approccio . . . . .                     | 14        |
| 4.5.1    | Modifiche alla libreria . . . . .                  | 14        |
| 4.5.2    | Localizzazione delle celle . . . . .               | 14        |
| 4.6      | L'approccio definitivo . . . . .                   | 15        |
| 4.7      | Risultati . . . . .                                | 17        |

|   |           |
|---|-----------|
| <b>5 Le modifiche al SIMIOR</b>                 | <b>19</b> |
| 5.1 Modifiche strutturali . . . . .             | 19        |
| 5.1.1 Il Database . . . . .                     | 19        |
| 5.1.2 L'Archivio statico . . . . .              | 20        |
| 5.2 Front-End . . . . .                         | 22        |
| 5.2.1 Collocazione della funzionalità . . . . . | 22        |
| 5.2.2 La schermata dei risultati . . . . .      | 23        |
| 5.2.3 Conferma dei risultati . . . . .          | 26        |
| <b>6 Conclusioni</b>                            | <b>27</b> |
| 6.1 Sviluppi futuri . . . . .                   | 27        |

# **Capitolo 1**

## **Introduzione**

Nelle Unità di Terapia Intensiva degli ospedali italiani viene effettuata, dal 2005, la sorveglianza delle Infezioni Correlate all'Assistenza (ICA) motivata dal fatto che i pazienti ricoverati nelle UTI presentano un rischio maggiore (dalle 5 alle 10 volte) di contrarre un ICA sia per fattori intrinseci che estrinseci, sia perché le UTI stesse sono epicentro di problemi emergenti di ICA. Oltre a questi due importanti motivi, la scelta di effettuare il monitoraggio è dovuta alla volontà di allinearsi ai progetti europei preesistenti e diventarne quindi componente della rete HAI-Net<sup>1</sup>, unendosi prima al network HELICS<sup>2</sup> e successivamente al progetto IPSE<sup>3</sup>.

### **1.1 SPIN-UTI**

L'obbiettivo del progetto SPIN-UTI (acronimo di *Sorveglianza attiva Prospettica delle Infezioni Nosocomiali nelle Unità di Terapia Intensiva*) è quello di assicurare la standardizzazione delle definizioni, della raccolta dei dati e delle procedure di feedback da parte delle strutture ospedaliere partecipanti alla sorveglianza nazionale ed europea delle ICA nelle UTI, col fine ultimo di migliorare la qualità dell'assistenza a livello europeo nelle UTI stesse.

---

<sup>1</sup>Healthcare-associated Infections Surveillance Network

<sup>2</sup>Hospital in Europe Link for Infection Control through Surveillance

<sup>3</sup>Improving Patient Safety in Europe

## 1.2 Soggetto di studio

Il reparto di terapia intensiva del Policlinico Ospedaliero Duilio Casula, al pari degli altri reparti nel sistema sanitario nazionale, adotta il sistema SPIN-UTI, con un implementazione informatica inadeguata per gli standard moderni. Per sopprimere a queste insufficienze la facoltà di Informatica dell'Università di Cagliari ha creato un nuovo sistema informatico denominato SIMIOR<sup>4</sup>, sviluppato su misura per le esigenze del reparto. I dati inseriti nel progetto contribuiscono al calcolo del punteggio di SPIN-UTI e le relative statistiche, ma la procedura di inserimento richiede attualmente un'operazione manuale da parte dei medici autorizzati. Ne conseguono dunque vari problemi, che spaziano dall'errore umano nella trascrizione alla pesantezza dell'operazione data dalla mole di dati da trascrivere. In questa tesi illustrato il lavoro e i punti cardine del lavoro effettuato per aggiungere al SIMIOR la funzionalità di analisi ed estrazione delle informazioni dai referti. Nelle prossime pagine verrà mostrata la situazione attuale di SPIN-UTI nell'ospedale, la soluzione proposta e i suoi dettagli, in particolare modo come sono costituiti i referti PDF e le tecnologie utilizzate per creare gli automatismi necessari a ricavarne informazioni, i problemi riscontrati durante il processo e i risultati ottenuti.

---

<sup>4</sup>Inserire acronimo simior

# **Capitolo 2**

## **SPIN-UTI e SIMIOR**

### **2.1 Implementazione di SPIN-UTI**

Il sistema SPIN-UTI 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. Nel mondo informatico, l'utilizzo accoppiato di questi due software è noto per essere un sistema debole di organizzazione e salvataggio dati, dettato dai limiti di archiviazione dello stesso Access, alla rappresentazione più criptica dei dati in Excel (soprattutto se confrontati con un'implementazione ad-hoc di un sistema di visualizzazione). Viene dunque logico, intuire l'inidoneità del sistema attuale, soprattutto sul fronte organizzativo dei dati.

### **2.2 Limitazioni del sistema attuale**

La principale limitazione dell'implementazione attuale del sistema implementato è il limite di informazioni (dette *record*) inseribili per ogni paziente, con la conseguente duplicazione delle schede 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. Anche la raccolta delle statistiche (fulcro del sistema SPIN-UTI) è pesantemente limitata da questa divisione, poiché il recupero delle stesse è reso difficoltoso. Infine, vi è il problema della manutenzione del sistema, difatti non vi è alcuna garanzia riguardo alla persistenza e alla sicurezza dei dati, ciò significa che non sono presenti sistemi che assicurino *backup* dei dati e protezione contro accessi malintenzionati.<sup>1</sup>

---

<sup>1</sup>Ad eccezione della semplice password a protezione dell'account utente, comunque insufficiente

## 2.3 Il progetto SIMIOR

Il Simior è un sistema informatico creato a inizio 2022 per sopperire alle limitazioni appena descritte, ed è attualmente utilizzato in fase sperimentale 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.

## 2.4 Differenze con l'implementazione attuale

Nel SIMIOR viene fatto un uso di un sistema di gestione di basi dati (*DMBS, Database Management System*) che permette tre vantaggi principali:

- Sopperire alle limitazioni di memoria di Access
- Costruire una struttura dati più efficiente e veloce nell'inserimento e nel recupero delle informazioni<sup>2</sup>
- Esprimere interrogazione più complesse

Visivamente, l'utente non viene caricato di tutte le informazioni del paziente scelto, ma ha subito un quadro chiaro del sistema con statistiche aggiornate automaticamente e la lista dei pazienti corredate di informazioni essenziali (es: data ricovero, codice paziente, motivo del ricovero). Solo una volta selezionato un paziente verranno mostrate più informazioni, sempre organizzate in relative sezioni, con eventuali grafici che meglio riescono a spiegare un determinato andamento. Oltre a questi vantaggi, il SIMIOR è modellato sulle esigenze specifiche del reparto e viene adattato di conseguenza all'evolversi delle necessità a differenza del binomio Excel-Access.<sup>3</sup> La funzionalità richiesta che ha portato alla stesura di questa tesi è l'inserimento dei referti di laboratorio per poterne estrarre informazioni, in particolare modo gli antibiogrammi, grazie ad essa è possibile automatizzare il processo di trascrizione dei referti, semplificando e velocizzando il lavoro dell'utente, con conseguente riduzione degli errori.

---

<sup>2</sup>in parte come beneficio ereditato dal tipo di tecnologia

<sup>3</sup>E' impossibile implementare funzionalità specifiche in questi software per via della loro natura *closed-source* che non li rende liberamente modificabili

# **Capitolo 3**

## **I referti medici**

### **3.1 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
- Più di pagina

### 3.1.1 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.



AZIENDA  
OSPEDALIERO  
UNIVERSITARIA  
DI CAGLIARI

Servizio Sanitario Regione Sardegna  
Azienda Ospedaliero Universitaria di Cagliari  
PP.OO. S.Giovanni di Dio - Policlinico D.Casula Monserrato  
U.O.S.C. Laboratorio Analisi Chimico Cliniche e Microbiologia  
Via Ospedale 54 - 09124 Cagliari  
Tel [REDACTED]  
Direttore: Dr. [REDACTED]

**Figura 3.1:** Intestazione iniziale di un referto

### 3.1.2 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.

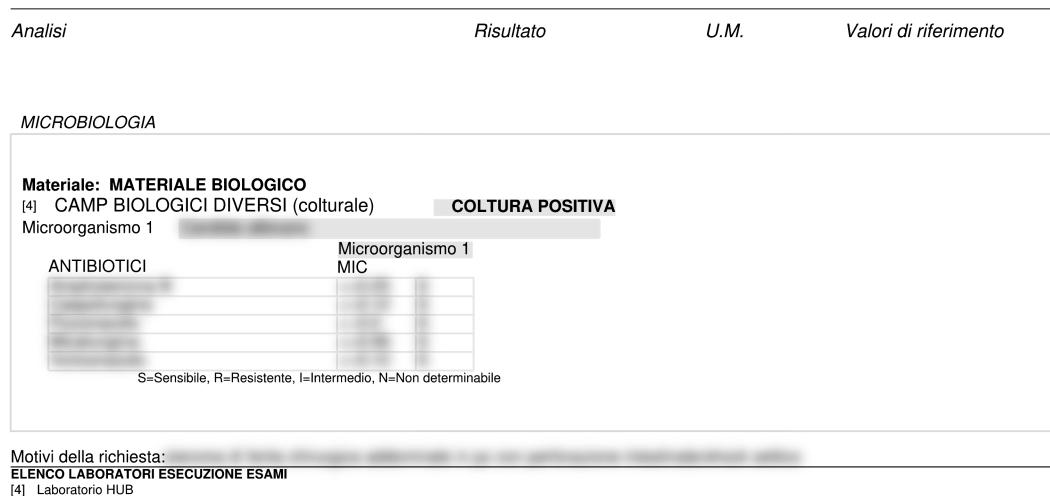
|                              |                 |                     |                          |                      |                   |
|------------------------------|-----------------|---------------------|--------------------------|----------------------|-------------------|
| Priorità: Urgenza            | Sig. [REDACTED] | Id.Paz.: [REDACTED] |                          |                      |                   |
| Richiesta: [REDACTED]        | del: [REDACTED] | Ore: [REDACTED]     | Data Nascita: [REDACTED] | Età: [REDACTED] Anni | Sesso: [REDACTED] |
| Provenienza: PO.TI BLOCCO TI |                 |                     | Cod.Fiscale: [REDACTED]  |                      |                   |
| Nosologico: [REDACTED]       |                 |                     |                          |                      |                   |

**Figura 3.2:** 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 così un'un'associazione erronea.

### 3.1.3 Il contenuto del referto

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



**Figura 3.3:** *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, denota 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  $\mu\text{g/mL}$ ) 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.

| <i>Analisi</i>  | <i>Risultato</i>  | <i>U.M.</i> | <i>Valori di riferimento</i> |
|---|---|-------------|------------------------------|
| <b>MICROBIOLOGIA</b>  |   |             |                              |
| <b>Materiale: BRONCO ASPIRATO</b><br>[4] BRONCOASPIRATO (Colturale) | <b>Risultato della coltura:</b><br>Assenza di crescita.   |             |                              |
| <b>Materiale: URINE</b><br>[4] URINOCOLTURA PER GG.CC. E MICETI     | <b>COLTURA NEGATIVA</b><br>(carica microbica < 1000 CFU/ml) per gram positivi, gram negativi e miceti. PAR test positivo. |             |                              |
| <b>ELENCO LABORATORI ESECUZIONE ESAMI</b><br>[4] Laboratorio HUB    |   |             |                              |

**Figura 3.4:** Un referto non valido per l'estrazione delle tabelle

Nella figura 3.4 viene mostrata la sezione contenuto di un referto dove non sono presenti tabelle estraibili, pertanto non valido ai fini delle funzionalità implementata.

### 3.1.4 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.

Pag.: 1 / 1

Referto firmato da Dott.: [REDACTED]

- Data di firma: [REDACTED]

Ore : [REDACTED]

**Figura 3.5:** Piè di pagina (Footer)

## 3.2 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 tavole

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 3.3 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 tavole multi-pagina costituiscono una variazione rara ma non impossibile delle prime due tipologie, presentando le tavole fisicamente divise fra una o più pagine.

### **3.3 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:

"Perché 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. Oltre tutto potrebbe essere necessario inserire referti non generati dallo specifico programma utilizzato in un determinato reparto.

# Capitolo 4

## 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.

### 4.1 Cosa sono i PDF?

PDF è un acronimo che sta per *Portable Document Format*, è un 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<sup>1</sup>, 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.

---

<sup>1</sup>Codice standard di codifica dei caratteri americano, uno standard *de-facto* utilizzato nei computer IBM

## **4.2 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.

## **4.3 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.

## **4.4 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 tabella tramite la seguente regex:

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

**Figura 4.1:** 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).

| Analisi   | Risultato  | U.M. | Valori di riferimento |
|---|--|------|-----------------------|
| <b>MICROBIOLOGIA</b>  |  |      |                       |
| <b>Materiale: BRONCO ASPIRATO</b><br>[4] BRONCOASPIRATO (Colturale) | <b>COLTURA POSITIVA</b>  |      |                       |
| Microorganismo 1<br>Microorganismo 2                                | Microorganismo 1 Microorganismo 2<br>MIC MIC                                 |      |                       |
| ANTIBIOTICI   |  |      |                       |
| S=Sensibile, R=Resistente, I=Intermedio, N=Non determinabile        |  |      |                       |
| <b>Materiale: URINE</b><br>[4] URINOCOLTURA PER GG.CC. E MICETI     | <b>COLTURA POSITIVA</b><br>Par test (Potere Antibatterico Residuo) positivo. |      |                       |

**Figura 4.2:** Tabella con più microrganismi

Per esempio, nella tabella mostrata nella figura 4.3 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 4.3:** Regex per l'estrazione delle righe della tabella

## 4.5 Il secondo approccio

Il secondo approccio, che apporta la modifica 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.

### 4.5.1 Modifiche alla 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.

### 4.5.2 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 dal 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.

## 4.6 L'approccio definitivo

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 4.3 dove avviene proprio l'esempio del materiale; infatti, la tabella relativa si trova nella seconda pagina (figura 4.4)

| <i>Analisi</i>  | <i>Risultato</i>  | <i>U.M.</i> | <i>Valori di riferimento</i> |
|---|---|-------------|------------------------------|
| <p>Microorganismo 1<br/>Microorganismo 2</p> <p>ANTIBIOTICI</p> <p>S=Sensibile, R=Resistente, I=Intermedio, N=Non determinabile</p> | <p>Microorganismo 1 Microorganismo 2</p> <p>MIC MIC</p> |             |                              |

---

**ELENCO LABORATORI ESECUZIONE ESAMI**  
[4] Laboratorio HUB

**Figura 4.4:**

La tabella in questione non presenta appunto il materiale su cui è basata l'analisi, poiché 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. 3.2)
  - (b) *Anagrafica* (fig. 3.3)
  - (c) *Piè di Pagina* (fig. 3.3)
2. Unire le pagine in una sola
3. Procedere dal punto 1) dell'algoritmo proposto nel paragrafo 4.5.2

## **4.7 Risultati**

I risultati sono stati raccolti da due fonti: le mail scambiate con i medici e i feedback rilasciati nella relativa sezione inserita nel progetto per questa tesi (vedi ??). Il sistema è stato giudicato ottimo sia per la semplicità che per la correttezza dell'implementazione, con una volontà da parte degli utenti di sfruttare il più possibile il sistema e trovare una soluzione agli impedimenti burocratici incontrati nella fase di distribuzione del progetto.



# Capitolo 5

## 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.

### 5.1 Modifiche strutturali

Le modifiche fatte si possono raggruppare in tre punti:

- Database
- Archivio Statico
- Front-End

#### 5.1.1 Il Database

Il database è stato alterato per permettere l'inserimento dei valori MIC (precedentemente il campo non esisteva e l'antibiogramma aveva solo la sensibilità per antibiotico), è stata creata inoltre la tabella feedback per permettere di raccogliere segnalazioni o eventuali richieste da parte degli utenti. La funzionalità di invio feedback è disponibile in ogni pagina del sito, accessibile tramite un click sul relativo tasto presente nella *navbar* di fianco al nome utente, la sua attivazione genera un *popup* con un campo di testo (di massimo 2048 caratteri) dove può essere inserito il messaggio. L'invio del feedback in questo modo, permette una raccolta più dettagliata delle informazioni utili al debug, quale la pagina su cui l'utente stava lavorando e i log del sistema.



**Figura 5.1:** *Navbar con sezione feedback e nome utente*

### 5.1.2 L'Archivio statico

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. Essi sono contenuti in un archivio, chiamato internamente *SimiorSelect*, da cui provengono tutti i valori mostrati negli elementi `<select>` del front-end (da cui il nome), qualsiasi operazione che richieda un inserimento di dati non numerici fa riferimento a questo archivio, e l'estrazione referti non è da meno. Il difetto della precedentemente implementazione riguarda la struttura di queste informazioni, difatti i nomi dei microrganismi e degli antibiotici e i loro codici erano distribuiti su due liste differenti, richiedendo diverse righe di codice ogni qualvolta fosse necessario associare le due informazioni, contribuendo all'errore umano in caso di modifica di queste informazioni (come l'aggiunta di ulteriori antibiotici). Si è proceduto quindi alla conversione di questo archivio dal formato XML (più ostico da leggere umanamente) a formato JSON e all'utilizzo di mappe nei casi più adatti (nell'esempio fatto prima). Seguono due esempi dell'organizzazione dei dati, prima (XML):

```
<antibiotico>
    <valore>Neomycin (oral)</valore>
    ...
    <valore>Nystatin</valore>
</antibiotico>
<cod_antibiotico>
    <valore>A07AA01</valore>
    ...
    <valore>A07AA02</valore>
</cod_antibiotico>
```

e dopo la conversione (JSON):

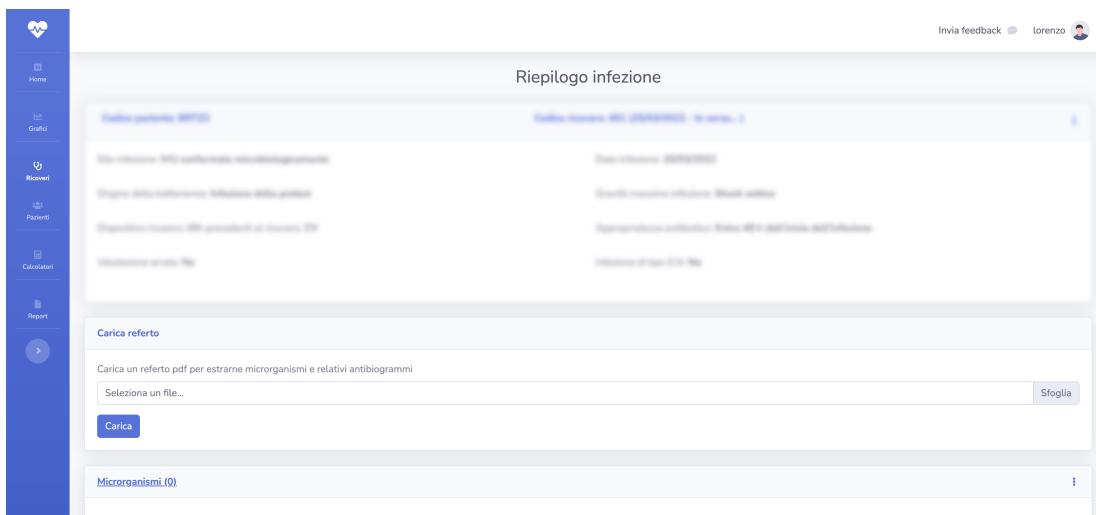
```
"antibiotici": {
    "A07AA01": "Neomicina (orale)",
    ...,
    "A07AA02": "Nistatina",
}
```

E' evidente la maggior semplicità (e leggibilità) della struttura utilizzata, che mostra i suoi effetti anche nelle prestazioni, infatti l'utilizzo di mappe Chiave-Valore permette un accesso più rapido rispetto a un doppio scorrimento della lista, difatti la complessità (in termini di tempo) della ricerca lineare è  $O(N)$ , visto lo scorrimento di due liste diventa  $2 * O(N)$ , l'accesso tramite mappa viene fatto calcolato l'hash che è sempre eseguito in tempo costante  $O(1)$ . Queste modifiche si rivelano particolarmente utili in tutti quei casi dove è necessario recuperare i codici degli antibiotici o dei microrganismi, come nel caso dell'inserimento di un antibiogramma (sia tramite la nuova funzionalità sia manualmente) dove è necessario convertire un nome *human-friendly* in una stringa identificativa univoca. Oltre alla conversione è stata effettuata anche la traduzione di determinati valori, difatti i referti indicano i antibiotici e microrganismi in italiano, mentre i dati dell'archivio derivano dal sistema internazionale di classificazione, pertanto sono indicati in lingua inglese. Esistono dei casi particolari in cui alcuni antibiotici indicati nei PDF hanno un nome alternativo, spesso il nome scientifico, che viene gestito con l'aggiunta della corretta alternativa fra parentesi vicino al nome principale (es: la *Benzilpenicillina benzatinica* comunemente conosciuta come *Benzatina penicillina G* sarà indicata come un solo nome così strutturato: "*Benzilpenicillina benzatinica (benzatina penicillina G)*")

## 5.2 Front-End

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. Ogni sezione è stata progettata per essere semplice e analoga ad altre parti del progetto per evitare confusione da parte dell'utilizzatore finale che è già abituato ai meccanismi del sistema. Infatti, la schermata dei risultati ha lo stesso aspetto dell'inserimento manuale dell'antibiogramma (una funzione di base del SIMIOR) con qualche aggiunta resa necessaria.

### 5.2.1 Collocazione della funzionalità



**Figura 5.2:** 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.

## 5.2.2 La schermata dei risultati

Una volta eseguita da parte del back-end l'elaborazione del documento, i risultati vengono mostrati all'utente il quale avrà la possibilità di apportare modifiche o confermare i dati estratti. Nella sezione superiore della pagina delle schede mostreranno un breve riepilogo, informando l'utilizzatore quali microrganismi hanno un antibiogramma valido ed eventuali microrganismi e antibiotici sconosciuti al sistema (quindi non presente nell'archivio statico, vedi sezione 5.1).

Risultato analisi referto

Microrganismi con antibiogramma valido :

| Antibiotico  | Sensibilità | MIC   |
|--------------|-------------|-------|
| Amoxicillina | • sensibile | ≤0,03 |

Conferma Annulla

**Figura 5.3:** Risultati estrazione

Se un microrganismo non è presente nel sistema non è possibile inserire neanche il relativo antibiogramma perché non sarebbe poi possibile farvi riferimento successivamente.



**Figura 5.4:** Microrganismo sconosciuto

Allo stesso modo vengono segnalati eventuali antibiotici sconosciuti, ma a differenza del caso precedente, è comunque possibile procedere con l'inserimento dell'antibiogramma che verrà mostrato mancante dell'antibiotico relativo.

I seguenti antibiogrammi contengono antibiotici che non sono stati inseriti perché non riconosciuti dal sistema

- Candida albicans
  - Amphotericina B (Mic: <=0,25, Sens: S (sensibile))

**Figura 5.5: Antibiotico sconosciuto**

In caso non tutti i dati siano estratti correttamente è possibile aggiungere o togliere manualmente informazioni, le opzioni di aggiunta sono disponibili nel menù drop-down posizionato in alto a destra nella scheda contenente la tabella. Per modificare un valore errato è sufficiente scegliere un'alternativa nella relativa lista, mentre per eliminare un valore (antibiotico o microrganismo) si seleziona l'elemento vuoto nella lista (indicato con un trattino), l'eliminazione ha un effetto "a cascata" che segue la seguente gerarchia: microrganismo → antibiotico → MIC e sensibilità, quindi l'eliminazione dell'antibiotico elimina l'intera riga mentre l'eliminazione del microrganismo elimina l'intero antibiogramma associato



**Figura 5.6: Drop-down con le opzioni**

Nel caso di più antibiogrammi estratti viene generata una pagina per ognuno

Risultato analisi referto

Microrganismi con antibiogramma valido :

- Micrococco resistente a ciprofloxacina-mecilamica
- Micrococco sensibile
- Enterococco Resistente
- Enterococco sensibile

Codice paziente: 00000000000000000000000000000000

Antibiogramma

Microrganismo

Antibiotico Sensibilità MIC

Indietro 1 2 3 4 Avanti

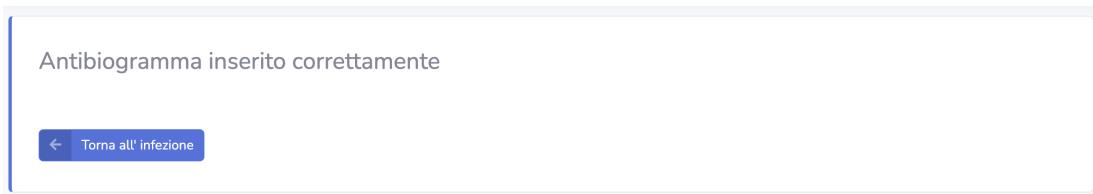
Conferma Annulla

Copyright © SIMIOR (UNICA) 2022-2023  
Per guida sull'uso: Leggi di più

**Figura 5.7:** Risultato estrazione di quattro antibiogrammi

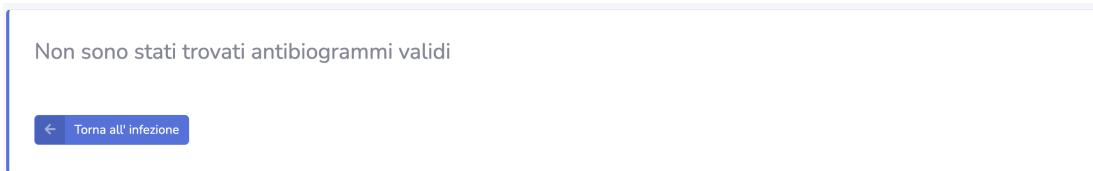
### 5.2.3 Conferma dei risultati

Una volta apportate le modifiche ritenute opportune, l'utente conferma i dati col relativo tasto, un messaggio indicherà il risultato dell'inserimento



**Figura 5.8:** Schermata di successo di inserimento

Se il referto non contiene alcuna informazione valida (come mostrato nella figura 3.4) viene mostrato direttamente un messaggio di avviso senza possibilità di inserimento, in caso si deve procedere con un inserimento manuale (disponibile nella pagina di riepilogo)



**Figura 5.9:** Messaggio di informazione

# **Capitolo 6**

## **Conclusioni**

In questa tesi abbiamo presentato l'implementazione della funzione di estrazione di dati clinici da referti, in un progetto nato per sopperire alle mancanze di un sistema adeguato nel reparto UTI del Policlinico, analizzando la situazione attuale nei suoi aspetti chiave con dettaglio, mostrando il progresso compiuto dal progetto nel suo primo anno di vita e illustrando la struttura dei documenti che sono stati il fulcro di questa ricerca. Questo studio ha portato alla contribuzione di una funzionalità chiave in un sistema robusto, capace di adattarsi alle esigenze più complesse con l'obbiettivo ultimo di migliorare la cura dei pazienti.

### **6.1 Sviluppi futuri**

L'estrazione delle tabelle antibiogramma è soltanto una parte dell'automazione implementabile tramite questo sistema, è infatti possibile espandere le funzioni per ottenere dai referti qualsiasi tipo di informazione. Sono attualmente implementati, ma non utilizzati, metodi per estrarre informazioni in modo mirato, generando un livello di astrazione utile per usi futuri che esulano il programmatore dal riscrivere parti complesse di codice. Un esempio sono le informazioni contenute nella sezione anagrafica (figura 3.2) per cui il sistema è già predisposto ad estrarne i dati, i quali permetterebbe la creazione di schede ricovero da un semplice referto. Queste limitazioni sono imposte da questioni di natura burocratica e legislativa, la prima impedisce un'espansione del sistema per via delle procedure richieste dalle aziende pubbliche in materia economica (es la gestione del sistema su cui il SIMIOR viene eseguito), la seconda impedisce una maggiore ricchezza di informazioni, sono infatti esclusi dettagli personali dei pazienti (es codice fiscale, nome e cognome).

