



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 2022/2023

In the working world, most digital documents are structured following the *de facto* standard called *Portable Document Format*, better known as PDF. Different environments ranging from schools to public administration, up to the national health service use this type of document to exchange information both with end users and between internal parties.

It is sometimes useful to be able to analyze these documents to extrapolate information, in this thesis two methods of data extraction will be analyzed and an implementation of a solution to this problem by analyzing the work done on the SIMIOR project.

Nel mondo lavorativo la maggior parte dei documenti digitali sono strutturati seguendo lo standard *de facto* chiamato *Portable Document Format*, meglio conosciuto come PDF. Ambienti diversi che variano dalla scuola all'amministrazione pubblica, fino al servizio sanitario nazionale utilizzano questa tipologia di documenti per scambiare informazioni sia con gli utenti finali sia fra le parti interne.

Risulta utile talvolta poter analizzare questi documenti per estrapolare informazioni, in questa tesi verranno analizzate due modalità di estrazione dei dati e un'implementazione di una soluzione a questo problema analizzando il lavoro svolto sul progetto SIMIOR.

Indice

1	Introduzione	1
1.1	Cenni storici	1
1.1.1	La struttura del PDF	1

Capitolo 1

Introduzione

1.1 Cenni storici

Il *Portable Document Format* è nato nel 1993 da Adobe Inc. con lo scopo di creare un formato file che permettesse di rappresentare documenti testuali o con immagini indipendentemente dall'hardware o dal software che usato per generare o visualizzare il documento stesso. Al momento la sua versione attuale è la 2.0 del 2017 ma lo standard più diffuso è la 1.4 di cui esiste la norma ISO 32000 stabilita nel 2008. Essendo un formato libero, possono essere creati programmi che leggono e scrivono documenti PDF senza la necessità di pagare diritti di utilizzo del brevetto. Nonostante il successo di questa tecnologia, il PDF non è considerato come adeguato alla conservazione sul lungo termine perché la riproduzione può dipendere da sorgenti esterne quali Font o oggetti esterni (link, immagini allegate esternamente ecc).

1.1.1 La struttura del PDF

