**Patent Related Information Conversion & Searching System**

Software Requirements Specification

Version 1.0

January 23, 2014

Himanshu Dogra

Lead Software Engineer,

Group 10

Prepared for

CS-258—Software Engineering

Instructor: Dr. Abhishek Srivastava

Spring 2015

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Description** | **Author** | **Comments** |
| Jan 23, 2014 | Version 1.0 | Himanshu Dogra | First Version of the Software |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Document Approval

The following Software Requirements Specification has been accepted and approved by the following:

|  |  |  |  |
| --- | --- | --- | --- |
| **Signature** | **Printed Name** | **Title** | **Date** |
|  | Himanshu Dogra | Lead Software Eng. |  |
|  | Dr. Abhishek Srivastava | Instructor, CS-258 |  |
|  | Dr. Ruchi Sharma | Client |  |

**Table of Contents**

[Revision History](#h.mcj8ym92zcr4)

[Document Approval](#h.w0auicnyv51y)

[1. Introduction](#h.1fob9te)

[1.1 Purpose](#h.dcu1kq6lbx4y)

[1.2 Scope](#h.pg27ld2ronuu)

[1.3 Definitions, Acronyms, and Abbreviations](#h.7y2dkxbc7ljx)

[1.4 Overview](#h.vmr4tg6us9n3)

[2. General Description](#h.owrznxce75i2)

[2.1 Product Perspective](#h.b3w4811pkl6c)

[2.2 Product Functions](#h.2bg0yppetpny)

[2.3 User Characteristics](#h.pfy9c7j66c6n)

[2.4 General Constraints](#h.fp5tw7mmyug)

[2.5 Assumptions and Dependencies](#h.15fd0yv0pt0s)

[3. Specific Requirements](#h.5tpiaigw8pyj)

[3.1 External Interface Requirements](#h.pbvg0lfnbz2j)

[3.1.1 User Interface](#h.l0orb1nkwhoj)

[3.1.3 Software Interface](#h.6x6l205govtc)

[3.1.4 Communications Interfaces](#h.jqtgn1o5cy96)

[3.2 Functional Requirements](#h.1y810tw)

[3.2.1 PDF data conversion into Tabular Format](#h.nb5zm113kv8k)

[3.2.2 Running a search query on the given tabular data](#h.2le6qkw22d07)

[3.3 Non-Functional Requirements](#h.7gt208kzz1ms)

[3.3.1 Performance](#h.xemxie4hmx9i)

[3.3.2 Reliability](#h.147n2zr)

[3.3.3 Availability](#h.3o7alnk)

[3.3.4 Security](#h.23ckvvd)

[3.3.5 Maintainability](#h.ihv636)

[3.3.6 Portability](#h.32hioqz)

# 

# 

# 

# 

# 1. Introduction

This is a complete software package that is used for the conversion of already existing data in .pdf file format into tabular data (.xls and/or csv file format). There is also a functionality to run a complex search query on the given tabular data with a search Algorithm running in the background. The accuracy and the speed of the conversion is the main target of the Algorithm.

## 1.1 Purpose

The purpose of this document is to present a detailed description of the Patent Data Conversion and Searching System. This document is intended for all the audiences who wish to contribute to this project and also for those who require a package with the functionalities existing in this software and/or want to add their own functionalities. It also explains what this software is capable of and what it is not capable of.

## 1.2 Scope

This software automates the process of conversion of existing patent related data in the pdf file format into tabular file format (.xls/ .csv). It will be a first software of its kind in India that does such conversion with great accuracy and speed on a large amount of data. It will also provide the functionality to run a complex search algorithm on the data that has already been converted into the tabular format. The search will be capable of running over a range of queries and will mainly be targeted towards the speed of fetching the results and the accuracy of the resulting data. The software will also be capable of communicating with the users through the Internet (via a Web Portal).

This software will only convert the files provided in a specific format. Any document provided in any other format may lead in less accurate results.

## 1.3 Definitions, Acronyms, and Abbreviations

**Algorithm:** The Algorithm used in this Software Package to search through the tabular data.

**Cloud:** Refers to anything over the Internet or the Internet itself.

**Conversion:** Refers to the conversion of patent related data from PDF file format to Excel and/or CSV file format.

**CSV:** Comma Separated Values, a common file format used to store tabular data.

**Graphical User Interface (GUI)**: The interface that allows the users to interact with the software through graphical icons, visual indicators, text, buttons etc.

**PDF:** A file format used to store data independent of the application software.

**Python:** The programming language used to code the software.

**Software Package/Application Software**: The complete set of package comprising of the actual software, the documentation to the software etc. which will enable a User to fully utilize its functionality.

**Searching:** Refers to the process of initiating the search algorithm on the given tabular data.

**User:** Reviewer or Author.

**Web Portal:** The online portal that can be used to access the software using the interval.

**XLS:** Another common file format used to store tabular data.

## 1.4 Overview

The next chapter, the General Description section, of this document gives an overview of the functionality of the product. It briefly describes the technicality required to efficiently analyze this product. The chapter follows with the constraints (or limitations) of the product and ending with the assumptions made in the implementation of this software.

The third chapter, Specific Requirements, is written primarily for the developers and describes the technical functionalities of the product.

Both the sections describe the same product in its entirety, but are intended for different audiences and thus use different language.

# 2. General Description

This software automates the process of conversion of existing patent related data in the pdf file format into tabular file format (.xls/ .csv). With the help of many search filters, we will be outputting the user-relevant data on another table which will make reading the relevant data easier. Till now there is no existing software for this. Presently, the only option to the user is to manually search the relevant data in each of the pdf files and filter it manually.

## 2.1 Product Perspective

The product is supposed to be an open source project hosted on GitHub. The product is supposed to be a self contained and independent package offering various functionalities. It provides a simple mechanism for users to efficiently manage their patents data.

The following are the main features that are included in the product.

* Automated File Conversion : Offers support for conversion of pdf tabular data into commonly known tabular data formats like csv, xls, etc.
* User account: The system allows the user to create their accounts in the system and provide features of updating and viewing profiles.
* Search: Search is simply local search engine based on key words and range queries.
* FAQs section: Frequently asked section contain answer of problem which users frequently face.

## 2.2 Product Functions

This product is aimed towards the automation of the process of conversion of the data related to the Patents (filed and accepted in India) provided in pdf file format into a tabular data format (xls/csv) so that it is easy to handle by the user. The functionality of running a complex search on the data in the tabular format will also be provided in the product. The product will also enables a user to access the above mentioned features from any Geographical Location via means of a web portal.

## 2.3 User Characteristics

Patent Citation Analyst, Intellectual Property Rights Researchers, Patent Policy Researchers and Intellectual Property Strategist.

## 2.4 General Constraints

* The main feature of the intended product is document conversion. But the conversion can take place only from the general .pdf format to .xls/.csv format.
* English is the only language throughout the implementation of this Software.
* The input needs to be in a specific format, violation of which may result in decreased accuracy.

## 2.5 Assumptions and Dependencies

The user must provide pdf format files which have specific layout, violation of which can lead to malfunctioning of the software.

# 3. Specific Requirements

# 3.1 External Interface Requirements

### 3.1.1 User Interface

This application shall provide a very intuitive and simple interface to the user and the administrator so that the user can easily navigate through conversion of data, searching for the required data and the administrator can easily manage groups and revoke user permissions.

**3.1.2 Hardware Interface**

**Recommended Requirements:**

|  |  |  |
| --- | --- | --- |
| **Processor** | **RAM** | **Disc Space** |
| To be estimated later | To be estimated later | less than 100MB (exact size to be determined later) |

### 

### 

### 3.1.3 Software Interface

* Client on INTERNET
  + Web Browser (any), Operating System (any)
* Client on a Local Machine
  + Conversion Software, Operating System (any)
* Data Base Server
  + SQL Server, Excel Server, Operating System (any)
* Development End
  + Python development kit (Python Libraries, Python GUI), HTML, XML, PHP clients, WEB SERVER (any)

### 3.1.4 Communications Interfaces

The HTTP or HTTPS protocol(s) will be used to facilitate communication between the client and the sever through the Internet. The software may communicate with the underlying database with the help of SQL queries.

## 3.2 Functional Requirements

### 3.2.1 PDF data conversion into Tabular Format

3.2.1.1 Introduction

One of the main purpose of this software will be the facility to convert a given Patent related data from pdf format into tabular format (xls/csv). The user will be provided with an interface in order to achieve it.

3.2.1.2 Inputs

User will select the pdf documents to be converted. The pdf file needs to be in a particular format.

3.2.1.3 Processing

An underlying Python program will process the given data depending on the keywords and subsequently store the results in a database.

3.2.1.4 Outputs

User will be provided with a xls/csv file containing the converted data.

3.2.1.5 Error Handling

The accuracy may be affected if a file of improper format is given in the input. Error handling will be done according to the latest standards.

### 3.2.2 Running a search query on the given tabular data

3.2.2.1 Introduction

Another main purpose of the software is to automate the searching process on the tabular data and quickly provide highly accurate results.

3.2.2.2 Inputs

User will input the search query in the form of keyword or ranges of queries over which the search is to be performed.

3.2.2.3 Processing

A searching algorithm will run on the given inputs. The speed and accuracy is targeted.

3.2.2.4 Outputs

User will be provided with another database containing the search results.

## 3.3 Non-Functional Requirements

### 3.3.1 Performance

The conversion will take very less time (less than a second for a single document). Due to high speed searching algorithm, 90% of the search queries can be answered in a second.

### 3.3.2 Reliability

The software aims at providing highly reliable and accurate results. The software code will take care of proper implementation of Error Handling Mechanisms. Many tests will be performed from time to time to ensure that the product is working properly with the addition of new features. An Algorithm will be specially implemented for the search purposes that will provide highly reliable and accurate search results.

### 3.3.3 Availability

The proposed site will be available for 99.99% time. The system downtime may not exceed 1 minute per day (depending upon the server hosting the website).

### 3.3.4 Security

Different users will have different permissions for the security reasons. Only the user with the proper authentication will be access the software online. Moreover, the database modifications and deletions on the local server will only be permitted to the administrator. Other users will be able to read the data depending upon their rights.

### 3.3.5 Maintainability

The database of the Patent related information may reside on a single or multiple systems, as required by the user. The database and the web-portal will be easily maintainable with the help of a very intuitive interface.

### 3.3.6 Portability

The software will be made such that it is easy even for the naive users. Moreover, the source will be provided on the Internet so that anyone can download and modify the software according to his/her requirements. The software will be accessible through a web portal so that anyone with an Internet connection can access the documents using their PC.