Table of Contents

Template Source:
http://web.cse.ohio-state.edu/~bair.41/616/Project/Example_Document/Req_Doc_Example.html

1. Introduction

1.1 Purpose of Document

This is a Requirements Specification document for a database containing information scraped from the Moody Manuals for the CU Boulder Leeds School of Business. The Leeds School of Business offers unparalleled access to world-changing entrepreneurs, national research institutions and award-winning faculty. Up to now, gathering data from these documents has been done manually, which has severely limited the amount of information that can be collected, and the universe of firms under study. The main goal of this project is to fully automate the data gathering process. This document describes the scope, objectives and goal of this project. In addition to describing non-functional requirements, this document models the functional requirements with use cases, interaction diagrams, and class models. This document is intended to direct the design and implementation of the database.

1.2 Project Summary

Project Name: Digitizing a Half-Century of Financial Panel-Level Data
Team Lead: Kevin Eastman
Project Manager: Qinglu Sun
NPL/ML Tech Lead: Jie Wang
OCR Tech Lead: Zuodong Wang
Database Tech Lead: Zijun Liu
Testing Lead: Joshua Khoo
Source Control Lead: Will Shanks
Documentation Lead: Remy Dahlke

1.3 Background

The goal of this project is to digitize and gather data from the Moody's manuals. These manuals, which are in the public domain prior to 1952, contain accounting, governance and other business information for listed corporations.

The Moody's manuals were initially published by John Moody starting in 1909 and have been published annually since then. For the period 1909 till the 1950s, the Moody's manuals are unrivaled as a source of business relevant information. We propose to scrape the data from the

Moody's Industrial, the Bank and Finance, the Transportation (Railroads), and the Public Utilities manuals.

Up to now, gathering data from these documents has been done manually, which has severely limited the amount of information that can be collected, and the universe of firms under study. The main goal of this project is to fully automate the data gathering process. The proposed project consists of two main parts: (i) Algorithmic method to extract historical data. To extract information from the Moody's manuals, the proposed project will do the following: (1) Create high quality digital images from each page of the Moody's manuals. (2) The resulting digital images will be converted to searchable text using a pre-process algorithm we have developed and existing OCR software. (3)We will extract information on the management and directors, credit ratings, geographical and other relevant business and financial information from the searchable text. Key project output from this part: At the completion of the project we will make our code opensource. A project of this scope should allow us to improve each stage of the algorithmic process and provide a clear template that future researchers can use for extracting data from other historical sources without relying on manual hand-collection that has been the standard up to this point. We note that the actual scanning of all images has already been done, but that there is the potential for taking more pictures of the microfiche using the scanner we have access to.

Currently, comprehensive historical firm-level US data do not exist prior to the 1970s. Lack of firm-level data makes it very hard to understand the dynamics of firms and business formation over time and through space as technology changes. While some researchers have hand-collected small datasets, no one has collected a large dataset of income statement and balance sheet information during the Great Depression, or the World War II period and the 1950s. The data generated by this project will have unprecedented breadths and scope and will provide firm-level data across industries, geography, and time. This broad scope of the micro-level data will allow significantly to improve our understanding of firm dynamics and business formation over much of the 20th century, as well as through space and technology evolution. Key project output from this part: The project will establish a database of firm-level business information from 1909-1952 from the Moody's Manuals 2. At the end of the project period this data will be posted on our public webpages, as well as on the ICPSR website next to the documentation detailing the extraction methodology and database guide.

The data extracted from the Moody's manuals will allow us to address a number of research questions that are impossible to answer with existing data. In particular, the data generated from this project will allow us to answer questions such as: what helped firms survive the Great Depression, cash/connections?; how did the Wild West become industrialized (what drove firms to Denver?); what was the effect of the Securities Act of 1933 on banks?; what effect did the introduction of the Securities and Exchange Commission (SEC) have on US companies?

1.4 Project Scope

In Scope:

- Extraction of Data
    - Create high quality post processed digital images from each page of the Moody's manuals.
    - Create a pre-process algorithm to get searchable text from our digital images.
    - Extract information on the management and directors, credit ratings, geographical and other relevant business and financial information from the searchable text.
- Database
    - Establish a database of firm-level business information from 1909-1952 from the Moody's Manuals
    - Implement a public UI.
- OCR
    - Tune the OCR to be optimal for the Moody Manuals.

Out of Scope:

- Any other business areas beyond those mentioned above
- Our database will not be searchable by images or contain images.
- We will not maintain the project after the Project/School Year is over.

1.5 System Purpose

1.5.1 Users

Those who will primarily benefit from the project are:

Students:

Upon completion of this project the students that are working on this project will have gained valuable experience in many fields of computer science as well as project management and teamwork skills. Students will become familiar with the practical implementation of coding language such as “R” and Python.

The teamwork and project management skills learned throughout the course of this project could possibly be the most valuable skills the students take away from it as they are vitally important to every workplace.

Researchers:

When we publish our database online researchers from all over the world will be able to access the data. This will be particularly interesting for researchers focusing on historical financial data.

An exciting part of this project is that our code will allow for other historical documents to be digitized for further study.

The Public:
The general public will also be able to benefit from this data as a possible point of interest. We will make available data on firms from the great depression era as well as who was involved with what companies during this time period.

1.5.2 Location
The database will be available to anyone using the Internet. Researchers will particularly benefit from the information that will become available that has been trapped in the paper of books for far too long.

1.5.3 Responsibilities
The responsibilities of this project are to:

- Provide a clear template that future researchers can use for extracting data from other historical sources without relying on manual hand-collection that has been the standard up to this point.
- Create high quality digital images from each page of the Moody's manuals that will be converted to searchable text using a pre-process algorithm we have developed and existing OCR software.
- Establish a database of firm-level business information from 1909-1952 from the Moody's Manuals

1.5.4 Need
The data generated by this project will have unprecedented breadths and scope and will provide firm-level data across industries, geography, and time. This broad scope of the micro-level data will allow significantly to improve our understanding of firm dynamics and business formation over much of the 20th century, as well as through space and technology evolution.

1.6 Overview of Document
The rest of this document gives the detailed specifications for the database. It is organized as follows:

Section 2: Functional Objectives
Each objective gives a desired behavior for the system, a business justification, and a measure to determine if the final system has successfully met the objective. These objectives are organized

by priority. In order for the new system to be considered successful, all high priority objectives must be met.

Section 3: Non-Functional Objectives (Jie Wang)
During making this project work, we will also build a more readable database, UI design and history system as non-functional objectives.

Section 4: Context Model
This section gives a text description of the goal of the system, and a pictorial description of the scope of the system in a context diagram. Those entities outside the system that interact with the system are described.

Section 5: Use Case Model
Use cases are a type of textual requirements specification that capture how a user will interact with a solution to achieve a specific goal. They describe the step by step process a user goes through to complete that goal using a software System. Use cases capture all the possible ways the user and system can interact that result in the user achieving the goal. They also capture all the things that can go wrong along the way that prevent the user from achieving the goal.

2. Functional Objectives

2.1 High Priority

The program shall create high quality post processed digital images from each page of the Moody's manuals.
The program shall create a pre-process algorithm to get searchable text from our digital images
The program shall extract information on the management and directors, credit ratings, geographical and other relevant business and financial information from the searchable text.
The database shall establish a publicly searchable database of firm-level business information from 1909-1952 from the Moody's Manuals.
The program shall tune the OCR to be optimal for the Moody Manuals.

2.2 Medium Priority

The system shall provide a search facility that will allow full-text searching of information on the management and directors, credit ratings, geographical,all relevant business and financial information that the user is permitted to access. The system must support the following searches: find all words specified

find any word specified
find the exact phrase

2.3 Low Priority

The program shall improve each stage of the algorithmic process including shorten the run time and improve accuracy for information extracted, provide a clear template that future researchers can use for extracting data from other historical sources without relying on manual hand-collection that has been the standard up to this point.

3. Non-Functional Objectives

3.1 Readable database: We are not only making it be searchable so that everyone can use that to search for text just based the scanned images. Also, the database should be clear that won't be confused by the users.

\

The way to achieve that: Have lines and columns with clear labels(describe what's the meaning of each)

3.2 UI design: The project should be used for everyone who is not familiar with the OCR. So we need to make this project be user friendly.

The way to achieve that: When searching the text through the engine, the engine will return the most related result. And then there are some buttons to report if there are some mistakes in the result.(feedback system)

3.3 History system: When users search for some words, the backend should memorize that. And users can review the search results in the future.

The way to achieve that: Build a database to record the search result. And link to each user id. And user id can be linked to Google account or other social media.

4. The Context Model

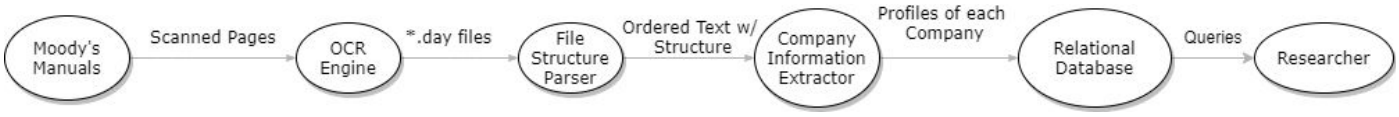
4.1 Goal Statement

The goal of the system is to digitize thousands of scanned pages from Moody's manuals from the first half of the twentieth century containing macroeconomic data and firm level financial data on listed companies. Currently, the scans are not helpful to researchers since they are simply images of text which is not searchable, so the project will make the manuals usable by:

- Using an OCR engine to determine raw text and location of the words in the scans

- Using clever parsing techniques to identify the structure of the pages—such as column, line, and chart delimitations—to inform the category of a section of text and what order the text should be read in
- Constructing standard profiles of each company by extracting consistent fields such as management and directors, credit ratings, geographical and other relevant business and financial information
- Making the information available in a searchable database
- Making the code open source

4.2 Context Diagram



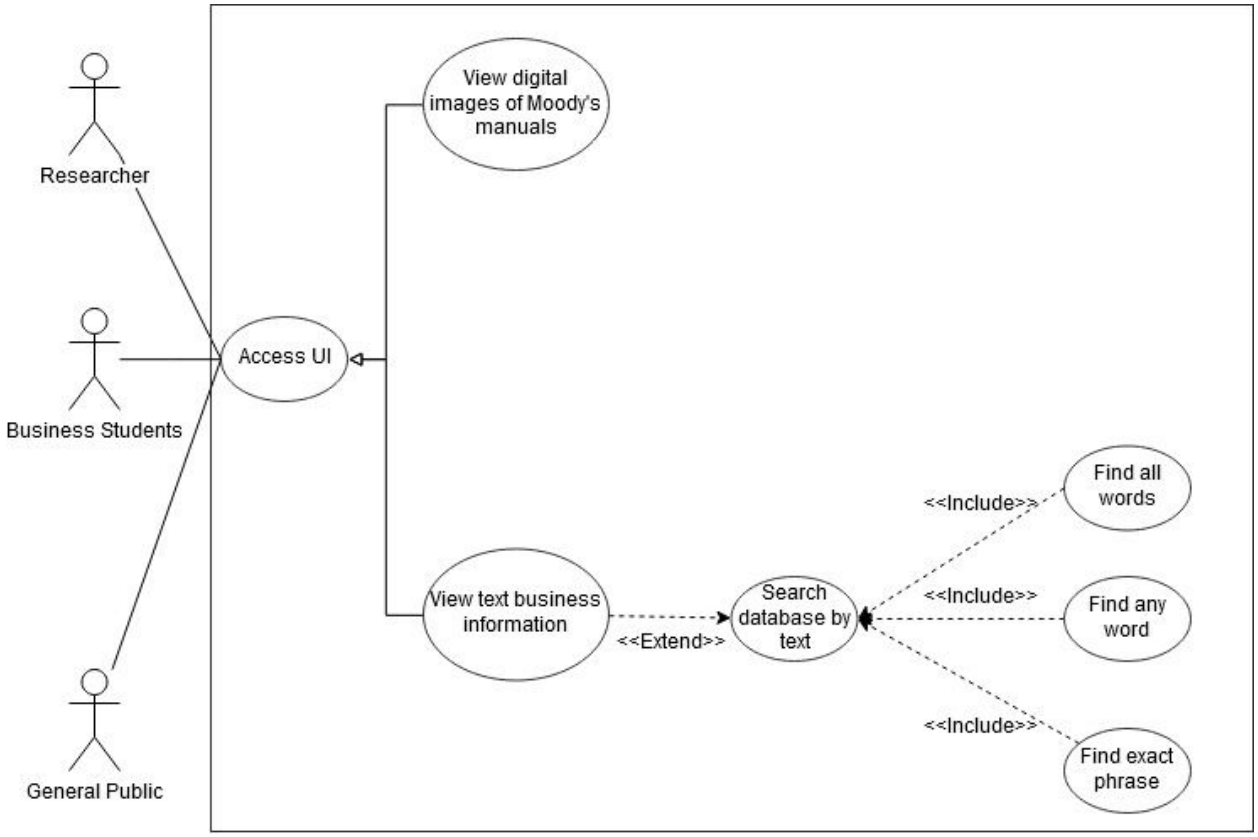
4.3 System Externals

Since the OCR Engine File, Structure parser, Company Information Extractor, Relational Database are all part of the method to our project, system externals will be the following:

- Potential researchers that do queries on our relational database and give feedback to the database.
- The moody's manuals that we scan to get page information and feed to OCR engine.

5. The Use Case Model

5.1 System Use Case Diagram



5.2 Use Case Descriptions (for selected cases)

Use Case Name:	Create / Update Database
Summary:	Using a script on the server the user is able to either add data to an existing database, or create a new one
Basic Flow:	1. The user uploads scanned images of the Moody's Manual pages they wish to add to the database 2. The user runs a script, supplying it with the names of the scans, and data

	about each page (what year/book/page it is) 3. The script creates .day files from running the scans through an OCR 4. The .day files are parsed and all relevant information is extracted 5. This information is added to the database
Alternative Flows:	Step 4: If the user supplies .day files instead, the script should be able to start at this step Step 5: If no database exists, a new one should be created If the new data is overwriting anything, the old data should be backed up for a period of time
Extension Points:	none
Preconditions:	User has scans of Moody Manuals
Postconditions:	New data is added to the database
Business Rules:	Must check if database exists, and create one if it does not Should sanity check new data, and backup any data being overwritten