

**“Semantic Documental Search”**

**A**

***Project Report***

*Submitted in partial fulfillment of the  
requirements for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE & ENGINEERING**

**With Specialization in Mainframe Technology**

**by**

<b>Name</b>	<b>Roll No.</b>
<b>Ankit Raj</b>	<b>R610213009</b>
<b>Mridu Gupta</b>	<b>R610213025</b>
<b>Neelesh Mehrotra</b>	<b>R610213030</b>
<b>Shivang Sharma</b>	<b>R610213048</b>

*under the guidance of*

**Mr. Pratyush Kumar Deka**

**Assistant Professor,  
CIT, UPES  
Dehradun**



**Department of Computer Science & Engineering**

**Centre for Information Technology**

**University of Petroleum & Energy Studies**

**Bidholi, Via Prem Nagar, Dehradun, UK**

**May-2016**



The innovation driven  
**E-School**

## **CANDIDATE'S DECLARATION**

We hereby certify that the project work entitled “**Automatic Reference Extractor**” in partial fulfillment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in Mainframe Technology and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of our work carried out during a period from **January, 2016** to **May, 2016** under the supervision of **Mr. Pratyush Kumar Deka, Assistant Professor, CIT.**

The matter presented in this project has not been submitted by me/ us for the award of any other degree of this or any other University.

**(Ankit Raj –R610213009)**

**(Mridu Gupta-R610213025)**

**(Neelesh Mehrotra- R610213030)**

**(Shivang Sharma- R610213048)**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 17/05/2016

**Mr. P.K Deka**  
Project Guide,  
Assistant Professor,  
CIT, UPES  
Dehradun

**Dr. Hanumat G Sastry**  
Program Head-Mainframe Technologies  
Center of Information Technology  
University of Petroleum & Energy Studies  
Dehradun-248001 (Uttarakhand)

## ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guides **Mr. Pratyush Kumar Deka** for all advice, encouragement and constant support they have given us throughout our project work. This work would not have been possible without their support and valuable suggestions.

We sincerely thank to our respected Program Head of the Department, **Dr. Hanumat G Sastry**, for his great support in doing our project in **Automatic Reference Extractor** at **CIT**.

We are also grateful to **Dr. Manish Prateek, Associate Dean CIT** and **Dr. Kamal Bansal, Dean COES, UPES** for giving us the necessary facilities to carry out our project work successfully.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

<u>Name</u>	<u>Roll No.</u>	<u>Signature</u>
Ankit Raj	R610213009	
Mridu Gupta	R610212025	
Neelesh Mehrotra	R610212030	
Shivang Sharma	R610213048	

## **ABSTRACT**

Search for scientific documents is a common and pertinent task that is faced by a lot of researchers as well as common Internet users in their daily searches. While going through a research paper one might want to refer to the scholarly articles cited in the reference section. Locating each of these articles manually over the web becomes a tedious task. The project deals with the mechanism of locating each of these papers over the web and determining whether they are accessible or not and if accessible, they will automatically be downloaded. What is taken into account is the title of individual reference string together with some additional metadata which forms our search expression that will be semantically searched over the web. We would implement a search algorithm which would make use of results fetched from Google Scholar for each search expression by means of web scraping.

**Keywords: semantic search, web scraping, Google Scholar.**

## Contents

<b><u>SNo.</u></b>	<b><u>Title</u></b>	<b><u>Page No.</u></b>
	<i>Certificate</i>	<i>ii</i>
	<i>Acknowledgement</i>	<i>iii</i>
	<i>Abstract</i>	<i>iv</i>
	<i>Contents</i>	<i>v</i>
	<i>List of Figures</i>	<i>viii</i>
<b>1.</b>	<b>Introduction</b>	<b>1</b>
1.1.	Overview	1
1.2.	History	2
1.3.	Problem Statement	2
1.4.	Motivation	2
1.5.	Objective	2
1.6.	Pert Chart Legend	2
<b>2.</b>	<b>System Analysis</b>	<b>3</b>
2.1.	Literature Review	3
2.2.	System Requirements	5
<b>3.</b>	<b>Design</b>	<b>6</b>
3.1.	Flow Chart Diagram	6
3.2.	Use-Case Diagrams	7
3.3.	Activity Diagrams	9
3.4.	Data Flow Diagram	10
3.5.	Process Model Used	11

<b>4. Algorithm</b>	<b>13</b>
<b>5. Screen Shots</b>	<b>18</b>
<b>6. Review</b>	<b>19</b>
6.1 Conclusion	19
6.2 Future Scope	19
<b>7. References</b>	<b>20</b>

## LIST OF FIGURES

<b>Fig. No. Name</b>	<b>Page No.</b>
<b>1. Introduction</b>	
Fig. 1.1. Pert Chart	2
<b>3. Design and Implementation</b>	
Fig. 3.1. Flow Chart	6
Fig 3.2. Use Case Diagram	7
Fig 3.3. Activity Diagram	9
Fig. 3.4.1. DFD Diagram: Dfd (Level 0)	10
Fig. 3.4.2. DFD Diagram: Dfd (Level 1)	10
Fig. 3.5. Process Model: Prototype Model	12
<b>5. Screen Shots</b>	
5.1 Start Screen	13
5.2 Analysis Screen	14
5.3 Extracted Metadata	15
5.4 Extracted Links	16
5.5 Download PDF	17
5.6 Opening Downloaded PDF	18

# 1. INTRODUCTION

## 1.1 OVERVIEW

A research paper is the culmination and the final product of an involved process of research, critical thinking, source evaluation, organization and the composition. A research paper tells us about the ongoing research in a particular field. Every research paper has its own reference section which tells us about the sources of information used in the text. It is possible for us to extract these references along with their metadata by using regular expressions. This citation metadata extracted forms a search expression which will serve as an input for semantic search. Our algorithm determines whether the source of the reference is available or not and then downloads all the available ones.

## 1.2 HISTORY

In the spring of 1884, a small group of individuals in the electrical professions met in New York, USA. They formed a new organization to support professionals in their nascent field and to aid them in their efforts to apply innovation for the betterment of humanity—the American Institute of Electrical Engineers, or AIEE for short.

A new industry arose, beginning with Guglielmo Marconi’s wireless telegraphy experiments in 1895-1896. What was originally called “wireless” telegraphy became radio with the electrical amplification possibilities inherent in the vacuum tubes that evolved from John Fleming’s diode and Lee de Forest’s triode.

Membership in both societies grew, but beginning in the 1940s, the IRE grew faster and in 1957 became the larger group. On 1 January 1963, the AIEE and the IRE merged to form the Institute of Electrical and Electronics Engineers, or IEEE.

By the early 21st century, IEEE served its members and their interests with 39 Societies; 130 journals, transactions, and magazines; more than 300 conferences annually; and 900 active standards and growing since then. With this number of scholarly articles hitting the internet, keeping a track of all these articles became a tedious task. This lead to the problem of locating these articles over the web that created the need of a “Reference Management Software” but none of them provided the functionality of locating these articles accurately and efficiently. Our **Automatic Reference Extractor** extracts the references from a research paper and helps us in locating these articles over the web.



### 1.3 PROBLEM STATEMENT

In the proposed system we shall be dealing with the problem of manually copying and pasting each reference cited in a research paper into a web browser to determine its source which is quite a tedious task given that there are dozens of references in a research paper. This results in wastage of time and involves too much of manual intervention.

### 1.4 MOTIVATION

We dealt with the problem of manually copying and pasting each reference cited in a research paper into a web browser to determine its source which is quite a tedious task given that there are dozens of references in a research paper. This results in wastage of time and involves too much of manual intervention. This motivated us to create an application that solves this problem and locates these articles automatically over the web.

### 1.5 OBJECTIVES

1. Making the search expression
2. Finding the source of the research article using web scraping.
3. Determining if the source is accessible or not.
4. Downloading the available sources.

### 1.6 PERT CHART LEGEND

A PERT chart is a project management tool used to schedule, organize, and coordinate tasks within a project.

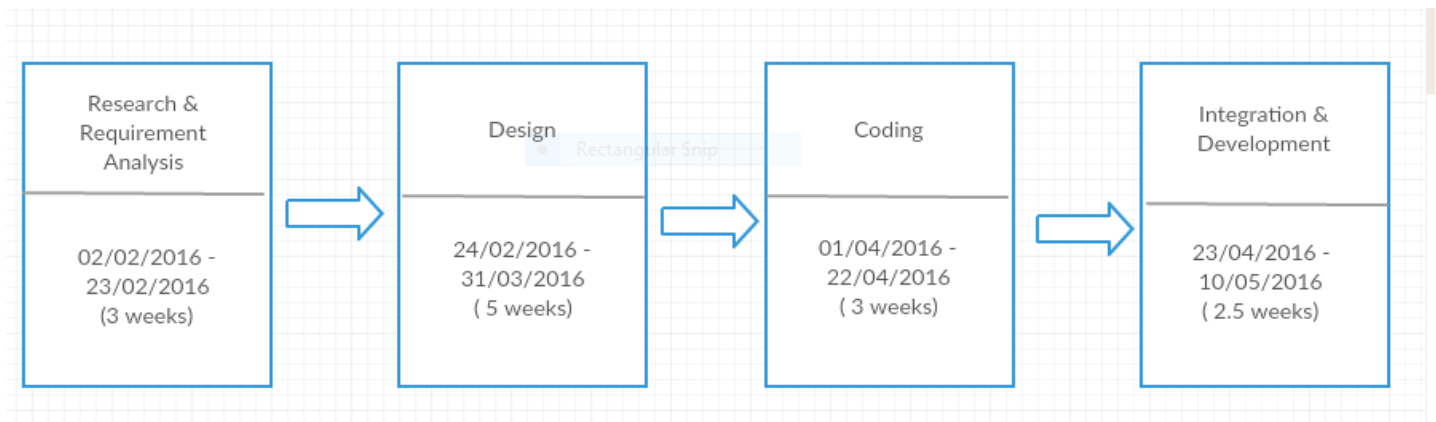


Figure 1.1: Pert Chart

## **2. SYSTEM ANALYSES**

### **2.1 LITERATURE REVIEW**

#### **Semantic Search**

Semantic search seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of the term as they appear in searchable data space, whether on the web or within a closed system to generate more relevant results. Google Scholar (a web crawler) semantically searches the web taking into account the documental context of the search expression

#### **Search Expression**

Each reference has its own metadata. This metadata gives us some additional information about the references. The reference metadata includes title, author name, volume, page number etc. This citation metadata forms our search expression with title being the most important part of it. This search expression is used as input by Google Scholar which fetches the most accurate results using its web crawling policy. Sara Pavia in their paper "A Fuzzy Algorithm for Optimizing Semantic Documental Searches" has talked about a strategy to make efficient search expression for semantic search [1].

#### **Web Scraping**

Web scraping (or web harvesting) is a computer software technique of extracting information from website's content. Since Google Scholar doesn't makes it content available via an API, we parse the results fetched by Google Scholar using web scraping. [2] We studied variety of open source HTML parsers to parse the web pages to search for valuable data embedded in the HTML page. One of the HTML parser is jsoup which provides efficient HTML parsing. [3] The applications that were quite helpful in our study were CiteSeerX, Google Scholar. These applications also uses an academic-focused crawlers to aid in there documental searches.

## **Google Scholar**

Google Scholar is a freely accessible web search engine that indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines.

We studied variety of open source web crawlers available in the market. There was a comparison between different open source and proprietary software along with their advantages and disadvantages. Some of them were Heritrix, Nutch, Scrapy, and YaCy.

## **Content Type**

A media type (also MIME type and content type) is a two-part identifier for file formats and format contents transmitted on the Internet. Before downloading the results from Google Scholar we have to identify its content type. Most of the research articles are available in the pdf format which can be downloaded directly. If the content type is html then we need to determine whether it's accessible or not.

## **JavaFX**

JavaFX is a software platform for creating and delivering desktop applications, as well as rich internet applications (RIAs) running across wide range of devices.

JavaFX Scene Builder is a visual layout tool that lets users quickly design Java FX application user interfaces, without coding. Users can drag and drop UI components to a work area, modify their properties, apply style sheets, and the FXML code for the layout that they were creating is automatically generated in the background. The result is an FXML file that can be combined with a java project by binding the UI to the application's logic.

## **2.2 SYSTEM REQUIREMENTS**

The project requires the following requirements for proper functioning:

- **Software Requirements:**
  - Windows 7.0\8.0
  - Java, JavaFx
  - Adobe PDF Reader.
  - Research PDF
- **Hardware Requirements:**
  - Processor Pentium IV and above
  - 512MB Ram
  - 25GB Hard Disk

### 3. DESIGN

#### 3.1 FLOW CHART DIAGRAM

A **flowchart** is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields.

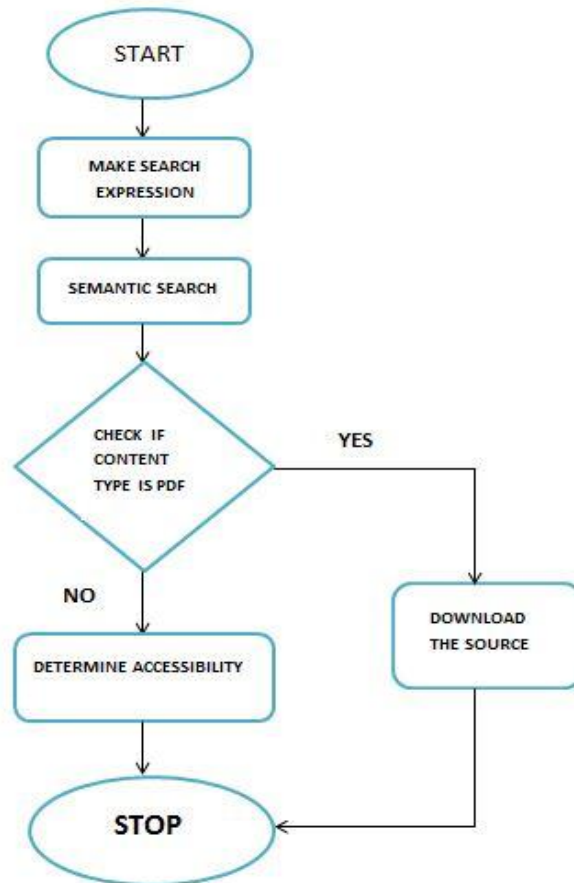


Figure 3.1: Flow Chart

### 3.2 USE CASE DIAGRAM [5]

A use case diagram is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases.

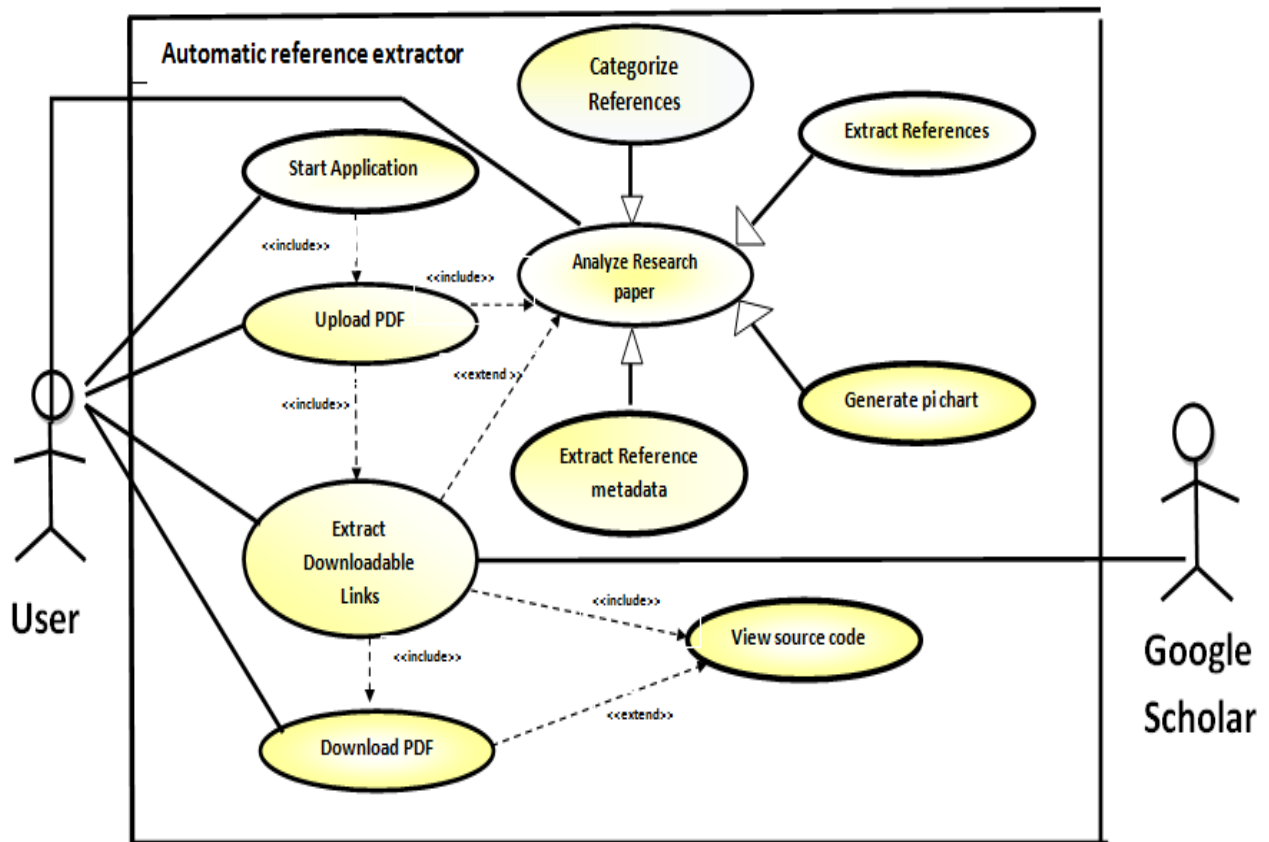
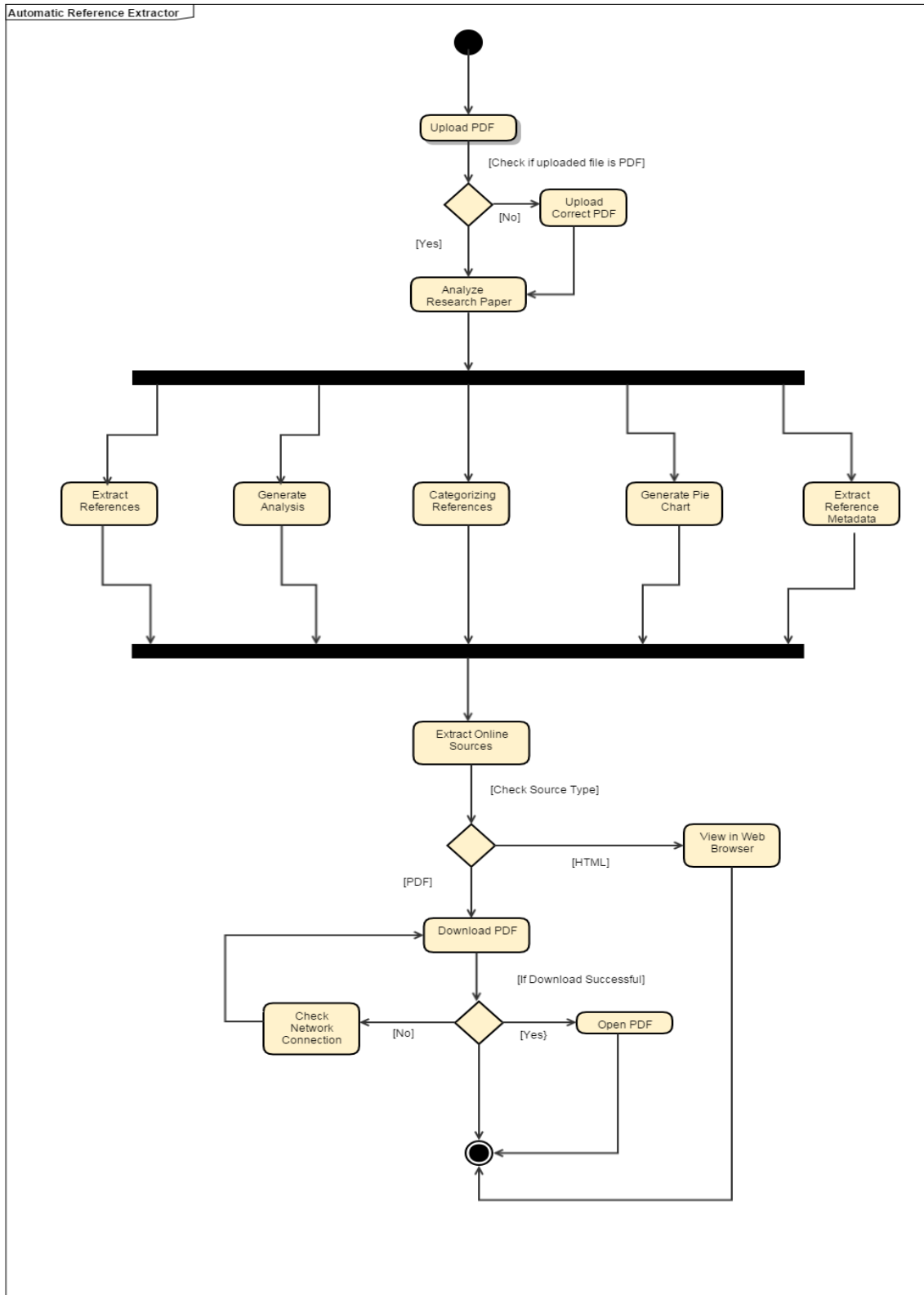


Figure 3.2: Use Case Diagram

## **3.2 ACTIVITY DIAGRAM**

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the unified modeling language, activity diagrams are intended to model both computational and organizational processes.



**Figure 3.3: Activity Diagram**



### 3.3 DATA FLOW DIAGRAM

Data flow diagram is an approach to visualize the data processing. A data flow diagram is strong in illustrating the relationship of processes, data stores and external entities in information system.

#### 3.3.1 DFD:LEVEL 0

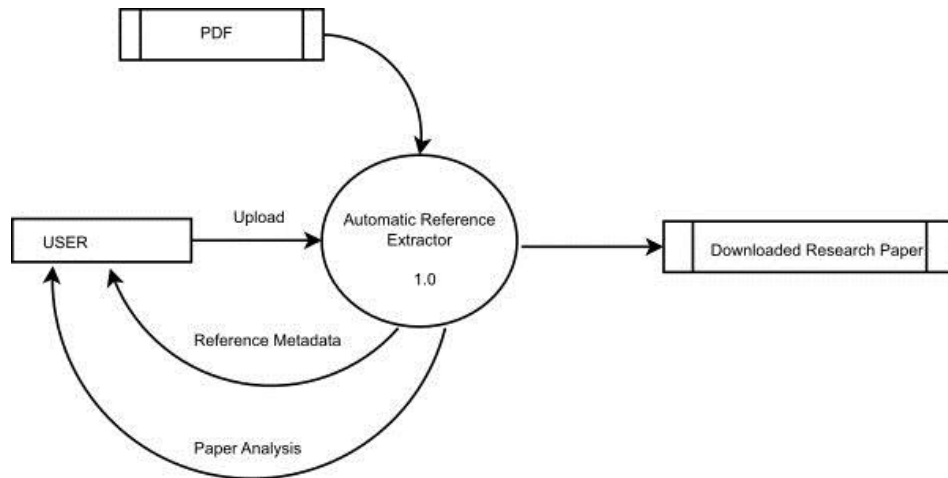


Figure 3.4.1: DFD (Level 0)

#### 3.4.2 DFD: Level 1

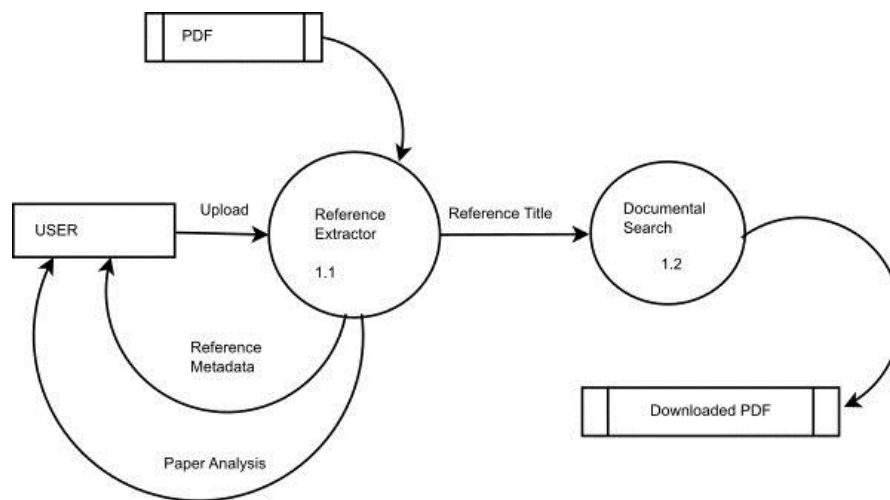


Figure 3.4.2: DFD (Level 1)

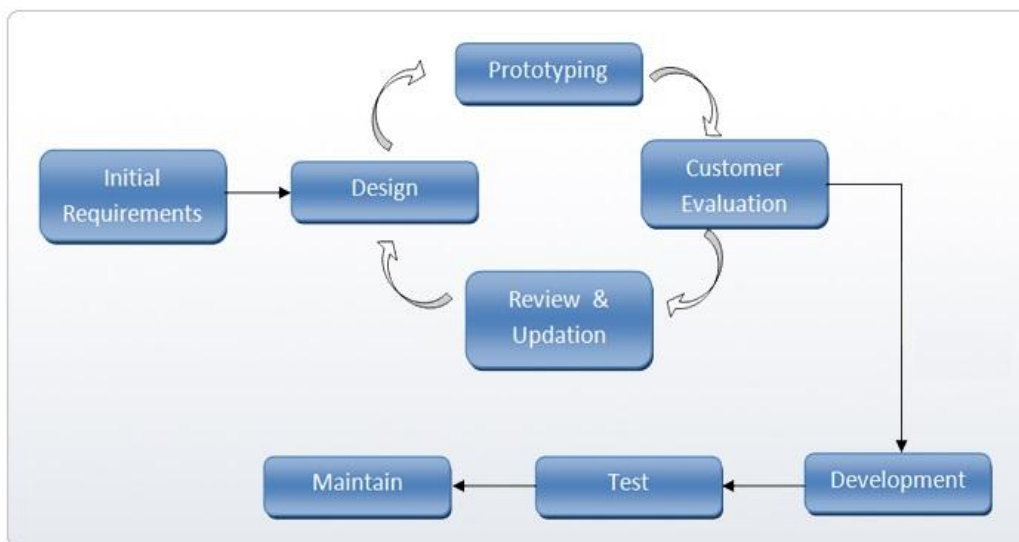
### 3.5 PROCESS MODEL USED

#### Prototype

Prototype is used as a project planning technique. This model suggests building a working prototype of the system, before development of the actual software. A prototype has limited functional capabilities, low reliability, or insufficient performance as compared to actual software. A prototype can be built very quickly by using several shortcuts. The shortcuts usually involve developing insufficient, inaccurate, or dummy functions.

Prototyping model is advantageous to use especially when exact technical solutions are unclear for development. A prototype can be helpful for developing the systems with unclear requirements and system with unresolved technical issues. Overall development cost can be turnout to be lower.

Prototypes are also advantageous to use for development of the graphical user interface (GUI) parts of the application. Through prototype it becomes easier to illustrate input data formats, messages, reports and the interactive dialogues to the user. For user it becomes much easier to form an opinion about the user interface, rather than imagining the working of a hypothetical interface.



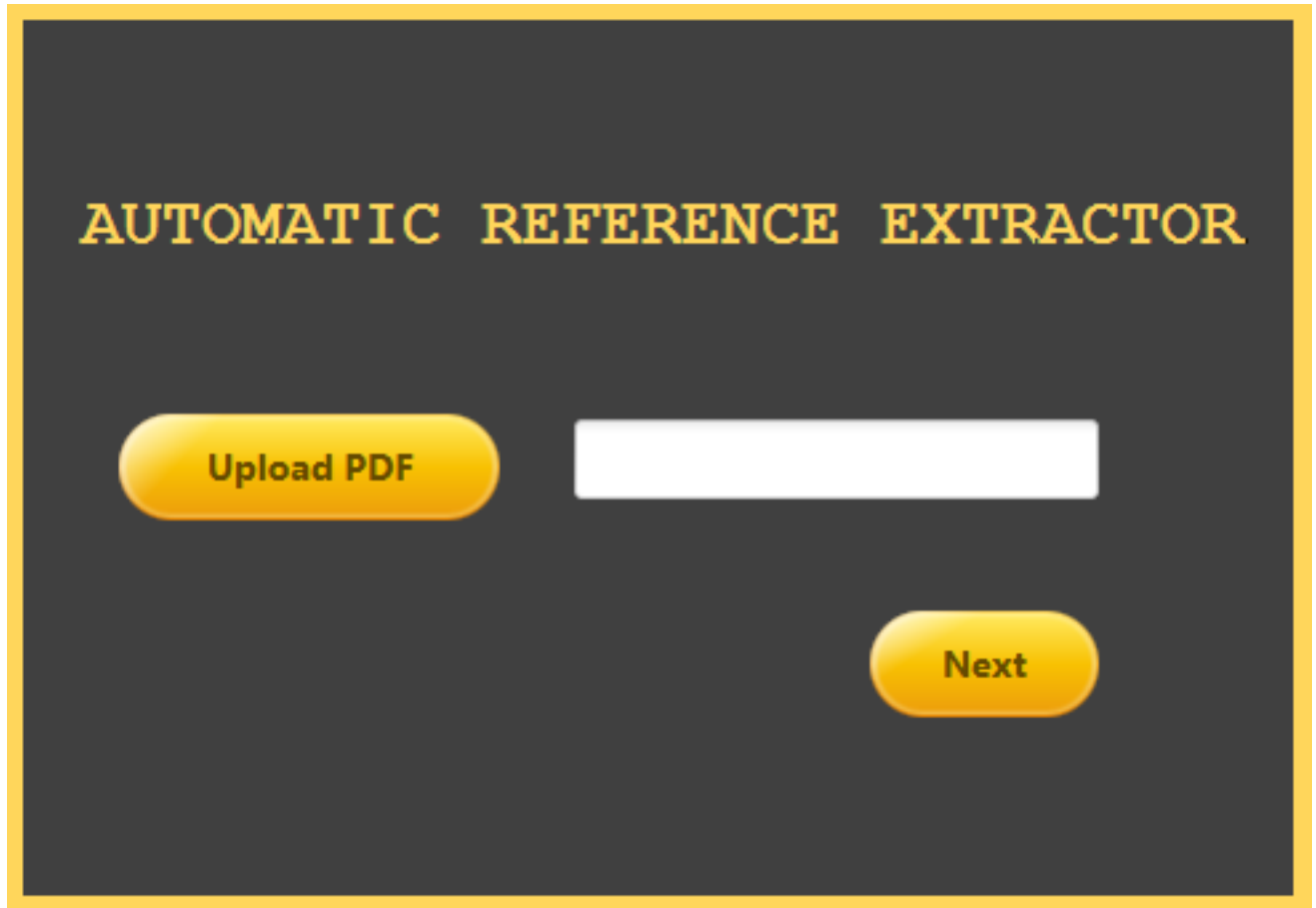
**Figure 3.5: Prototype Model**

## 4. ALGORITHM

- Step 1. Start
- Step 2. Take PDF file as input.
- Step 3. Extract reference section using regular expression\*
- Step 4. Store extracted section in a variable OutputText.
- Step 5. Split OutputText into individual reference strings using regular expression\*\*
- Step 6. Store individual references in an array.
- Step 7. For each value of array DO
- Step 8. Apply regular expression to know the type of reference.
- Step 9. IF value matches the regular expression
- Step 10. GOTO Step 12.
- Step 11. Else GOTO Step 8.
- Step 12. Apply regular expressions\*\*\* to extract citation metadata.
- Step 13. Apply regular expression to extract citation title.
- Step 14. Store Extracted Title in a Variable SearchExpression.
- Step 15. Extract Links for SearchExpression using JSoup.
- Step 16. For each link determine Content Type.
- Step 16. IF Content Type=PDF
- Step 17. Download PDF.
- Step 18. Else Open link in Web Browser.
- Step 19. End If.
- Step 20. End For
- Step 21. End If.
- Step 22. End For
- Step 23. End

## 5. SCREENSHOTS

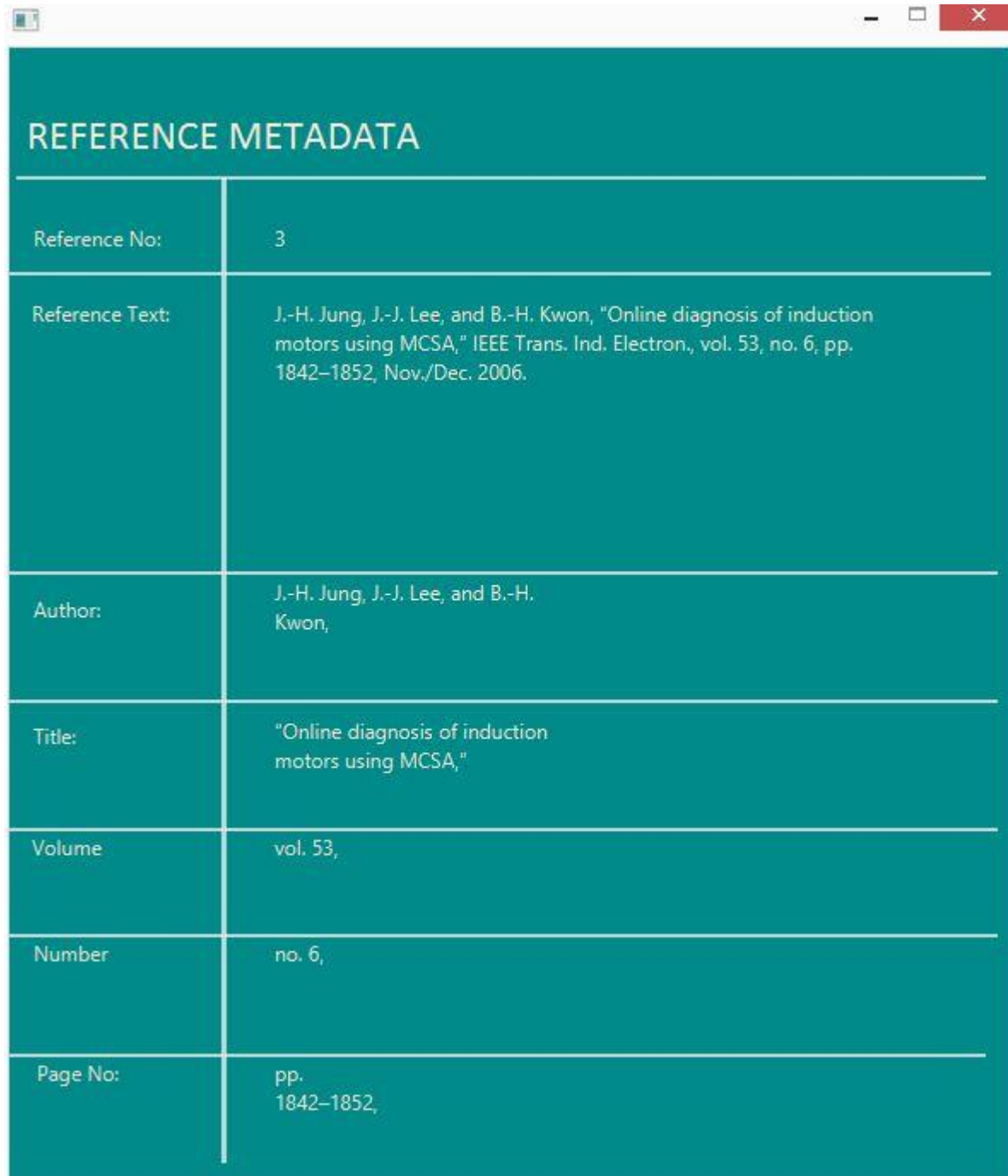
### 5.1 Start Screen



## 5.2 Analysis Screen



### 5.3 Reference Metadata



REFERENCE METADATA	
Reference No:	3
Reference Text:	J.-H. Jung, J.-J. Lee, and B.-H. Kwon, "Online diagnosis of induction motors using MCSA," IEEE Trans. Ind. Electron., vol. 53, no. 6, pp. 1842–1852, Nov./Dec. 2006.
Author:	J.-H. Jung, J.-J. Lee, and B.-H. Kwon,
Title:	"Online diagnosis of induction motors using MCSA,"
Volume	vol. 53,
Number	no. 6,
Page No:	pp. 1842–1852,

## 5.4 Extract Links

The screenshot displays a web browser interface. On the left, a sidebar titled "No of links found 2" lists two links under "All Links found":  
1. <https://www.smp.uq.edu.au/people/YoniNazarathy/Control4406/...>  
2. <http://www.sciencedirect.com/science/article/pii/S000510980100...>  
Below the list, there are fields for "Html Link" and "PDF Link", both pointing to the second link. A "Download Status" section shows a "Done" checkbox. An "Open PDF" button is located at the bottom of the sidebar.

The main browser window shows the ScienceDirect website. The header includes the ScienceDirect logo and navigation links for "Journals" and "Books". Below the header, there are "Purchase" and "Export" buttons. The main content area features the Elsevier logo and the journal title "Automatica", along with the issue information: "Volume 38, Issue 1, January 2002, Pages 3–20". The article title "The explicit linear quadratic regulator for constrained systems" is prominently displayed, with a star icon below it.

At the bottom of the browser window, a "References" section is visible, listing several related works:

- A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, "The explicit linear quadratic regulator for constrained systems," *Automatica*, vol. 38, no. 1, pp. 3–20, Jan. 2002.
- S. J. Wright, "Applying new optimization algorithms to model predictive control," *Chemical Process Control-V*, vol. 93, no. 316, pp. 147–155, 1997.
- F. A. Potra and S. J. Wright, "Interior-point methods," *J. Comput. Appl. Math.*, vol. 124, no. 1–2, pp. 281–302, 2000.
- E. A. Yildirim and S. J. Wright, "Warm-start strategies in interior-point methods for linear programming," *SIAM J. Opt.*, vol. 12, no. 3, pp. 782–810, 2002.
- S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Eng. Practice*, vol. 11, no. 7, pp. 733–764, 2003.
- W. Wang, D. E. Rivera, and K. Kempf, "Model predictive control strategies for supply chain management in semiconductor manufac-

## 5.5 Download PDF

The screenshot displays a web application interface for downloading a PDF. On the left, a sidebar shows search results with the following details:

- No of links found:** 2
- All Links found:**
  - <https://www.smp.uq.edu.au/people/YoniNazarathy/Control4406/...>
  - <http://www.sciencedirect.com/science/article/pii/S000510980100...>
- Html Link:** <http://www.sciencedirect.com/science/article/...>
- PDF Link:** <https://www.smp.uq.edu.au/people/YoniNa...>
- Download Status:**  Done
- Open PDF** button

The main content area shows a preview of the article 'Automatica' from ScienceDirect. The article title is 'Automatica', Volume 38, Issue 1, January 2002, Pages 3–20. The Elsevier logo is visible. Below the article preview, there is a red banner that says 'Important Message' and a note 'Powered by DNSUnlocker'.

At the bottom of the interface, there is a 'References' section with the following list of references:

- A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, "The explicit linear quadratic regulator for constrained systems," *Automatica*, vol. 38, no. 1, pp. 3–20, Jan. 2002.
- S. J. Wright, "Applying new optimization algorithms to model predictive control," *Chemical Process Control-V*, vol. 93, no. 316, pp. 147–155, 1997.
- F. A. Potra and S. J. Wright, "Interior-point methods," *J. Comput. Appl. Math.*, vol. 124, no. 1–2, pp. 281–302, 2000.
- E. A. Yildirim and S. J. Wright, "Warm-start strategies in interior-point methods for linear programming," *SIAM J. Opt.*, vol. 12, no. 3, pp. 782–810, 2002.
- S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Eng. Practice*, vol. 11, no. 7, pp. 733–764, 2003.
- W. Wang, D. E. Rivera, and K. Kempf, "Model predictive control strategies for supply chain management in semiconductor manufac-



## 5.6 Opening Downloaded PDF

1.pdf - Adobe Reader

File Edit View Document Tools Window Help

1 / 18 61.8% Find

Pages

1 2 3 4

PERGAMON

Automatica 38 (2002) 3-20

www.elsevier.com/locate/automatica

automatica

### The explicit linear quadratic regulator for constrained systems<sup>☆</sup>

Alberto Bemporad<sup>a,b,\*</sup>, Manfred Morari<sup>b</sup>, Vivek Dua<sup>c</sup>, Efstratios N. Pistikopoulos<sup>c</sup>

<sup>a</sup>Dip. Ingegneria dell'Informazione, Università di Siena, Via Roma 56, 53100 Siena, Italy  
<sup>b</sup>Automatic Control Laboratory, ETH Zentrum, ETH 120, 8092 Zurich, Switzerland  
<sup>c</sup>Centre for Process Systems Engineering, Imperial College, London SW7 2BZ, UK

Received 24 September 1999; revised 9 October 2000; received in final form 16 June 2001

*We present a technique to compute the explicit state-feedback solution to both the finite and infinite horizon linear quadratic optimal control problem subject to state and input constraints. We show that this closed form solution is piecewise linear and continuous. As a practical consequence of the result, constrained linear quadratic regulation becomes attractive also for systems with high sampling rates, as on-line quadratic programming solvers are no more required for the implementation.*

**Abstract**

For discrete-time linear time invariant systems with constraints on inputs and states, we develop an algorithm to determine explicitly, the state feedback control law which minimizes a quadratic performance criterion. We show that the control law is piece-wise linear and continuous for both the finite horizon problem (model predictive control) and the usual infinite time measure (constrained linear quadratic regulation). Thus, the on-line control computation reduces to the simple evaluation of an explicitly defined piecewise linear function. By computing the inherent underlying controller structure, we also solve the equivalent of the Hamilton-Jacobi-Bellman equation for discrete-time linear constrained systems. Control based on on-line optimization has long been recognized as a superior alternative for constrained systems. The technique proposed in this paper is attractive for a wide range of practical problems where the computational complexity of on-line optimization is prohibitive. It also provides an insight into the structure underlying optimization-based controllers. © 2001 Elsevier Science Ltd. All rights reserved.

**Keywords:** Piecewise linear controllers; Linear quadratic regulators; Constraints; Predictive control

### 1. Introduction

As we extend the class of system descriptions beyond achieved with a non-linear control law. The most popular approaches for designing non-linear controllers for linear systems with constraints fall into two categories:

## 6. REVIEW

### 6.1 CONCLUSION

A research paper is the culmination and the final product of an involved process of research, critical thinking, source evaluation, organization and the composition. A research paper tells us about the ongoing research in a particular field. Every research paper has its own reference section which tells us about the sources of information used in the text.

Search for scientific documents is a common and pertinent task that is faced by a lot of researchers as well as common Internet users in their daily searches. While going through a research paper one might want to refer to the scholarly articles cited in the reference section. Locating each of these articles manually over the web becomes a tedious task.

With the help of our **Automatic Reference Extractor**, a researcher can extract reference metadata, get an insight about the type of reference, and generate a paper analysis report. Moreover using the **Automatic Reference Extractor**, user does not have to search the web for individual references. A user can depend on this product, which in turn will search the web for a reference and also provide its accessibility information. If there are PDF links found, they are automatically downloaded on the user machine. This solves the problem of manual work done to locate each reference individually over the web.

### 6.2 FUTURE SCOPE

- This application can be added as an extension to the already existing reference management softwares such as Mendeley, Zotero, Endnote etc.
- This application can also be used for different publications like IEEE, Springer, Elsevier etc.

## 7. REFERENCES

[1] Sara Paiva. a fuzzy algorithm for optimizing semantic documental searches. web. source available : <http://www.sciencedirect.com/science/article/pii/S2212017313001552>.

[2] "Web Scraping". Wikipedia.n.p.,2016.web.17 mar. 2016.

[3] "jsoup". oracle. oracle foundation, n.d. web. 20 aug. 2015. source available at: <http://jsoup.org/>.

[4] "Google Scholar". wikipedia.n.p.,2016.web.17 mar. 2016.

[5]Lillian Rostad. 2012. An extended misuse case notation: Including vulnerabilities and the insider threat Vol. 1: 1-11.