# Clippy: Smart PDF Reader for Better Paper Reading Experience and Knowledge Mining

**Group Number:** 2022604
**Supervisor Name:** Shurui Zhou
**Student Names:**
Priscilla Deng - 1005227826
Bethany Chu - 1005277111
Marie Joy Cuevas - 1005408989
Hannah Ruiz - 1005408961

## 1.0 Executive Summary

Academic papers are often lengthy, but necessarily so, as they contain an abundance of information. As reading these papers is a tedious but essential task for researchers, we plan to reduce reading and comprehension time by creating a more intuitive PDF Reader catered towards academia use. The PDF Reader will include three main features: a quick way to view cross-references (figures and tables), a citation knowledge graph and research summaries with highlighted key phrases in the PDF document. Using PDF.js as the base, the PDF reader will incorporate Semantic Scholar's TLDR to create automatic summaries and use an external JS Library to create knowledge graphs for references of the paper. With these functions, the created PDF Reader should reduce researchers' reading and comprehension time and inspire new research ideas.

## 2.0 Motivation

There are existing PDF readers, such as Hammer PDF and Paperly, that assist researchers in understanding academic papers. However, neither PDF readers are able to implement knowledge mining to provide additional insight to the paper, for example, in the form of knowledge graphs and summaries [1,2]. Furthermore, Hammer PDF only provides a quick jump feature when clicking on cross-references and is unable to display the referenced data alongside the document [1]. While Paperly can cross-reference citations through a pop-up overlay when mousing over the reference in the document, it is unable to do so for other data types, such as figures and tables [2]. Our PDF reader, Clippy, will resolve these missing features to provide the researcher with a more comprehensive reading experience.

## 3.0 Problem Statement

Researchers take hundreds of hours every year to read academic papers [3]. We plan to create an interactive PDF reader for academic documents, designed to improve a researcher's reading experience.

## 4.0 Project Goal

To create a smart PDF reader making an academic paper more comprehensible for researchers.

## 5.0 Scope of Work

The PDF reader will create features based on PDF.js to avoid reimplementing basic parsing features [4]. The additional features will include a citation knowledge graph, highlighting key phrases and quick cross-referencing. The project will utilize the Semantics TLDR's database to find summaries and key phrases of the academic paper [4]. A graph generator will also be implemented with the design to create the knowledge graph feature. Figure 1 shows the different tools that the project will be using to create a solution.

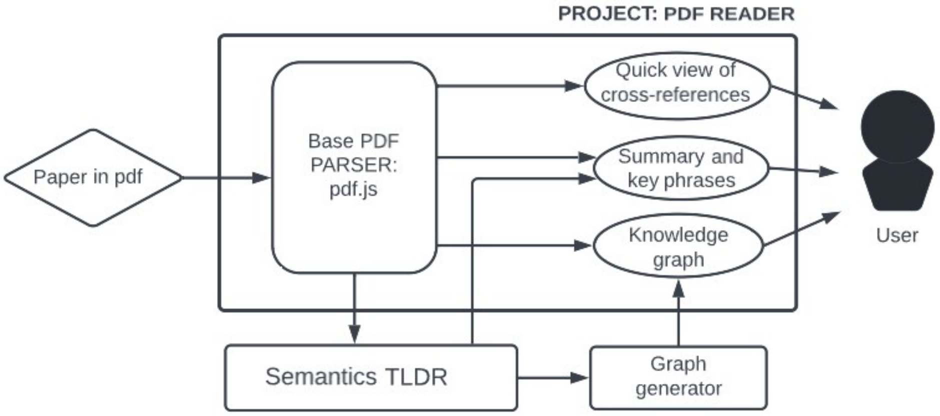

Figure 1: The workflow of available tools used for the solution's design

### 6.0 Requirements

For the creation of an interactive PDF reader, the following functions, objectives, and constraints are to be considered when implementing the project.

### 6.1 Functions

The design will create an interactive PDF reader for academic documents, with general PDF reader functionalities and specific functions related to improving comprehension time. The following functions are divided into primary functions, the central functions of the design, and secondary functions, functions that enable or result from the operation of a primary function.

#### 6.1.1. Primary Functions

- Hovering over cross references (figures, tables and citations) allows the reference to be displayed without changing the current page
- Displaying the knowledge graph for all citations/references
- Presenting the summary of PDF document

#### 6.1.2 Secondary Functions

- Extracting data from a PDF file
- Building the knowledge graph
- Summarizing the contents of the academic paper
- Highlighting key sentences used to form the summary

### 6.2 Objectives

The design will have various objectives that outline what should be implemented in the solution. The objectives are listed in descending order of importance.

**Table 6.2.1** Objectives, Metrics, and Goals

| Objective | Metric | Goal |
|---|---|---|
| Summary should be concise | Number of words in summary | < ⅓ original text [5] |
| Quick summary and knowledge graph loading time | Time needed to load the summary and knowledge graph | < 2 seconds [6] |
| Quick cross reference loading time | Time needed to load the cross reference | < 1 second [2] |
| Make PDF document more comprehensible | Comprehension test score obtained within a predetermined amount of time | > 70%  [7] |

### 6.3 Constraints

The constraints are strict requirements on what the design must include. There are multiple physical client requirements, as well as functional requirements.

- Cross references, summary, and knowledge graph must load within 10 seconds [8]
- Must use PDF.js library [4]
- Must use Semantic Scholar API for summarizing [4]
- MIT license [4]
- No matter the size of the displays, PDF should still be readable. Minimum font size is 8 points in Times New Roman for readability [9]

### 7.0 Conclusion

To optimize a researcher's reading experience, we propose a smart and interactive PDF reader capable of displaying pop-ups containing cross-reference data efficiently, creating knowledge graphs to show the relationship between cited papers, and generating summaries to give a brief overview of the academic paper. These features should reduce the researchers' reading and comprehension time and also inspire new research ideas. Further research is required to understand how to test levels of reading comprehension in higher education.

## 8.0 References

[1] S. Wang et al., "Hammer PDF: An Intelligent PDF Reader for Scientific Papers", 2022. [Online]. Available: https://arxiv.org/pdf/2204.02809.pdf. [Accessed: 19- Sep- 2022].

[2] "Paperly—A paper reader designed for researchers", *Medium*, 2019. [Online]. Available: https://medium.com/paperly/paperly-a-paper-reader-designed-for-researchers-dc6b3af34817. [Accessed: 18- Sep- 2022].

[3] S. Keshav, "How to Read a Paper", *Web.stanford.edu*. [Online]. Available: https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf. [Accessed: 19- Sep- 2022].

[4] X. Xie, Y. Li and Y. Tang, "Clippy: smart PDF reader for better paper reading experience and knowledge mining". [Online]. Available: https://docs.google.com/document/d/1cfioPoOR8dD9vLzjQ0hPULgdjKeNBKVyWZJ8vxE5GSo/edit. [Accessed: 14- Sep- 2022].

[5] J. Buckley, *Fit to print: the canadian student's guide to essay writing. 2nd ed*. Toronto: Harcourt Brace Jovanovich, 1995.

[6] C. Tran, "Extractive Summarization with BERT", *Chris Tran*, 2020. [Online]. Available: https://chriskhanhtran.github.io/posts/extractive-summarization-with-bert/. [Accessed: 19- Sep- 2022].

[7] J.-R. Wang, S.-F. Chen, R.-F. Tsay, C.-T. Chou, S.-W. Lin, and H.-L. Kao, “Developing a test for assessing elementary students' comprehension of science texts - international journal of science and mathematics education,” *SpringerLink*, 29-Jun-2011. [Online]. Available: https://link.springer.com/article/10.1007/s10763-011-9307-y. [Accessed: 18-Sep-2022].

[8] J. Nielsen, "Response Time Limits: Article by Jakob Nielsen", *Nielsen Norman Group*, 1993. [Online]. Available: https://www.nngroup.com/articles/response-times-3-important-limits/. [Accessed: 17- Sep- 2022].

[9] H. AKUTSU, "Character Legibility 1 : Character Size and Legibility Evaluation", 2008. [Online]. Available: https://www.jstage.jst.go.jp/article/jjsse/12/2-2/12_94/_article/-char/ja/. [Accessed: 19- Sep- 2022].