AIO: File Type Conversion

Submitted by

Ansh Sachdeva (41325502716)
Nitin Kumar (41725502716)
Kunal Wadhawan (01225502716)

Submitted to

Mr. Pankaj Singh Yadav (Project guide)

Department of Computer Science and Engineering in

Partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY

Department of Computer Science and Engineering

JIMS Engineering Management Technical Campus, Greater Noida

Guru Gobind Singh Indraprastha University, Delhi

MAY, 2020

CERTIFICATE

This is to certify that Project Report entitled “AIO: File Type Conversion” which is submitted by Ansh Sachdeva, Kunal Wadhawan, Nitin Kumar in partial fulfilment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of JIMS Engineering Management Technical Campus, Greater Noida affiliated to Guru Gobind Singh Indraprastha University, Delhi is a record of the candidate own work carried out by him under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date: 01/05/2020 Mr. PANAKJ SINGH YADAV

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature
Ansh Sachdeva (41325502716)

Signature

Kunal Wadhawan(01225502716)

Signature

Nitin Kumar (41725502716)

Approved By:

Signature	**Signature**
Mr. Pankaj Singh Yadav	**Dr. Vikas Chaudhary**
(Project Guide)	**(HOD)**
CSE Department	CSE Department
JIMS EMTC, Greater Noida	JIMS EMTC, Greater Noida

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Mr. Pankaj Singh Yadav, Department of Computer Science & Engineering, JIMS Engineering Management Technical Campus, Greater Noida for his constant support and guidance throughout the course of our work.

His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavours have seen light of the day. We also take the opportunity to acknowledge the contribution of Professor (Dr.) Vikas Chaudhary Head, Department of Computer Science & Engineering, JIMS Engineering Management Technical Campus, Greater Noida for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members and Lab staff of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature
Ansh Sachdeva (41325502716)

Signature

Kunal Wadhawan(01225502716)

Signature
Nitin Kumar (41725502716

TABLE OF CONTENTS

ABSTRACT

During the last two decades, the use of internet has been changing every domain of technology. It has also led to the tremendous development and implementation of software using web technologies from the last few years. There are lots of platform for converting file types online but has data privacy issue, and some software offline but has size limitation, platform requirements. Therefore, AIO: File Type Conversion is only one platform or PWA (Progressive Web App), which is small in size, work completely offline, and platform independent (only web browser required) to convert various file types to respective formats. There will be no usage of server other than first-time website launch in browser. This platform only uses open-source framework and library, and some self-created combinational algorithms to perform file type conversion.

Keywords: open format; offline PWA, propriety format,

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

PWA	Progressive Web App
JPEG	Joint Photographic experts' group
GIF	Graphics Interchange Format
PNG	Portable Network Graphics
PDF	Portable Document Format
SVG	Scalable Vector Graphics
MD	Mark Down
TXT	Plain text
UTF	Unicode Transformation Format
ASCII	American Standard Code for Information Interchange

Chapter 1

Introduction

1.1 Background

A Computer File is a blob of information stored on a memory device. These files are used for recording data discretely and are managed and organized by a File system.
Data conversion is the process of conversion of computer data from one format to another. We often need to convert a file to a different format based on our needs. Throughout a computer environment, data is encoded in a variety of ways. For example, computer hardware is built on the basis of certain standards, which requires that data contains, for example, parity bit checks. Similarly, the operating system is predicated on certain standards for data and file handling. Furthermore, each computer program handles data in a different manner. Whenever any one of these variables is changed, data must be converted in some way before it can be used by a different computer, operating system or program. Even different versions of these elements usually involve different data structures. For example, the changing of bits from one format to another, usually for the purpose of application interoperability or of capability of using new features, is merely a data conversion. Data conversions may be as simple as the conversion of a text file from one-character encoding system to another; ormore complex, such as the conversion of office file formats, or the conversion of image formats and audio file formats.

There are many ways in which data is converted within the computer environment. This may be seamless, as in the case of upgrading to a newer version of a computer program. Alternatively, the conversion may require processing by the use of a special conversion program, or it may involve a complex process of going through intermediary stages, or involving complex "exporting" and "importing" procedures, which may include converting to and from a tab-delimited or comma-separated text file. In some cases, a program may recognize several data file formats at the data input stage and then is also capable of storing the output data in a number of different formats. Such a program may be used to convert a file format. If the source format or target format is not recognized, then at times a third program may be available which permits the conversion to an intermediate format, which can then be reformatted using the first program. There are many possible scenarios.

Successful data conversion requires thorough knowledge of the workings of both source and target formats. In the case where the specification of a format is unknown, reverse engineering will be needed to carry out conversion. Reverse engineering can achieve close approximation of the original specifications, but errors and missing features can still result.

- **Proper Planning:** Successful data conversion requires proper planning. Any data conversion project needs to start with defining the boundaries of the project. You can plan your project by asking a variety of questions to define these boundaries. These include:

    - What kind of data needs to be converted?
    - What is the quality of data and its availability? Does it require full or partial conversion?
    - Which data should be moved to the new database?
    - Which data should not be moved?
    - What kinds of formats are needed for data conversion? For example, your project may require SGML, XML, HTML, and other formats. An efficient SGML conversion outsourcing provider may execute this task and deliver the product quickly. You can then use this data for your SGML-compatible databases.
    - What is the original data format and what is the final format?
    - What would be the extent of digitization?
    - Is the destination database compatible with the software used for data and HTML conversion?
    - What are the data conversion standards to be used, if any, for the successful completion of data conversion projects?
    - What are the guidelines for the process?
    - What would be the tentative duration of project?
    - How frequently would do you need to carry out the data conversion?
    - The cost of data conversion is one of the limiting factors for a data conversion project.

- **Ensure Business Engagement:** Sometimes, those in the business side of the organization do not understand the importance of data conversion. It is important to make them understand the importance of data conversion in data processing and management: about how the quality of data conversion can affect subsequent processes that makes use of this data. Data conversion is thus, a task critical from both business and technical perspectives.

- **Implement Data Standards:** Defining and implementing data quality standards helps to ensure consistency across the different databases. Consistently measure and track data quality and constantly check the effect on the business value. (hyperlink to article on data quality impact on business value).

- **Data Profiling and Cleansing:** Ensure that proper data profiling and data cleansing procedures are in place so that the original data is of high quality. This helps to smoothen out the subsequent data conversion procedures.

- **Data Management and Data Governance:** Following data conversion, ensure that the duplicate master data is eliminated, reducing the risk of incorrect transactions and unreliable reports. The project should satisfy all principles of data management and data governance.

But there are many challenges and principals involved in the conversion of file from one format to another:

- **Simple or Complex**: A conversion can be as simple as converting a text file from one-character encoding system to another (for e.g. converting a markdown file to html or converting a colored image to grayscale). It can also be as complex as conversion between different image formats, which requires deeper understanding of compression and encoding algorithms.

- **Seamless or processing involvement: A** conversion can be as seamless as upgrading a file to a newer version of a computer program (like converting a WordStar 2 file (ws2) to WordStar 3 (ws3)). Alternatively, a conversion can require processing by the use of a special conversion program.

- **Single or multiple stages**: A conversion could happen in a single stage or it may involve a complex process of going through intermediary stages.

- **Export/Import procedures:** A conversion may involve complex "exporting" and "importing" procedures.

1.2. Problem Statement

During the last two decades, the use of internet has been changing every domain of technology. It has also led to the tremendous development and implementation of software using web technologies from the last few years. There are lots of platform for converting file types online but has data privacy issue, and some software offline but has size limitation, platform requirements. Therefore, AIO: File Type Conversion is only one platform or PWA (Progressive Web App), which is small in size, work completely offline, and platform independent (only web browser required) to convert various file types to respective formats. There will be no usage of server other than first-time website launch in browser. This platform only uses open-source framework and library, and some self-created combinational algorithms to perform file type conversion.

Objective: Creating a secure and free platform for the file format conversion which supports document, audio, image conversion along with the added feature for the file compression. This platform use client-side javascript for file conversion it does not going to store users file on anywhere on server. Server will use only to host web application for anywhere access. All the Computation will be done at client side by user machine only.

1.3. Purpose and Goals

The main purpose of the project is to implement a platform (*aio.io*) for converting files from one format to another **without the use of any server.** Various image (png, jpeg, svg, gif), video (mp4, flv), audio (mp3, ogg) and file formats (docx, pdf, txt, html, etc.) would be converted interoperability between the same file type.

Features:

- **Completely offline**: Files are converted and stored on the user device only
- **Free to use**
- **Better Security**: No logs are created on the Website Server
- **Faster file conversions:** Since user's device is being used to perform conversions, we get incredible speeds without compromising on efficiency
- **Small Application Size**

This project came to existence because of there is no platform available for file type conversion in small application, work without internet as most of online application are small in size but they store user private files to their server in for converting file, which make user privacy to risk. That's why human need a platform for such purpose.

Chapter 2

Literature Survey

---

1. A General Methodology for Data Conversion and Restructuring

Publisher: IEEE, Author: V.Y. Lum, N.C. Shu, B.C. Housel
Published in: IBM Journal of Research and Development (Volume: 20 , Issue: 5 , Sep. 1976)

Abstract:

This paper presents a methodology and a model for data conversion or translation. The model assumes that both source and target systems are available and that conversion interfaces may be required to interact between these systems and the conversion system. To achieve data conversion or translation using this approach, two languages are needed: 1) a language to describe the data structures, and 2) a language to specify the mapping between source and target data. This paper describes these two languages, DEFINE and CONVERT and gives numerous examples to show the capabilities of these languages and how they can be used in data conversion and restructuring. Both languages are high level and nonprocedural and have the power to deal with most situations encountered in data conversion processes. In addition, the paper also describes some of the facilities in the languages specifically designed for data checking in a data conversion process.

2. A Comparative Study of Data Migration Techniques

Publisher: Ajeet Ghodeswat, Trupti Shah, Amruta Mhatre, Santosh Dodamani
Published in: IOSR Journal of Engineering (IOSRJEN), Volume 9, PP 77-82

Abstract:

This paper gives the overview of the data migration and its basic concepts. The need for the data migration in terms of business needs, different environments and the criticality of the legacy databases are discussed. The data migration process generally carried out in three steps: Plan, Migrate, and Validation. An automated ETL (Extract-Transform-Load) Process/Tools is use to move the data from source to target database. ETL is the process of data cleaning, transforming and finally loading it into a new target database. Existing Methodologies Like Line of Code, Sampling Techniques and MINUS Queries Are Briefly Discussed in The Paper. During the Process of The Data Migration (DM), Too Much Data Is Extracted, Transformed, Structured, And Loaded (ETL) From Legacy/Source Database into A Newer Structure That Is the Target Database. This Process Leads to Various Types of Changes In Data, Known As Data Inconsistencies /Quality Issues.

3. Decoding and Conversion of TFD Logging Data Format Based on Java

Publisher: Mingxing Liang, Peipei Zhang
Published in: International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2015)

Abstract:

Matrix logging system is widely used in coal-bed methane (CBM) reservoir. TFD is the extension of Matrix output file which is a structured format comprised a mix of ASCII and binary. Well CAD is the only logging software which can be used because of the confidentiality format, so that the processing and interpretation results of CBM reservoir are less effective. This paper decodes the TFD format and analyzes the storage pattern. Based on this, Java programming language is used to develop a format conversion program. The application can accurately convert the TFD format data to LAS (Log ASCII Standard), which can be loaded into the majority of well logging software platform. Result shows that the description of DAT structure and outcome of DAT file conversation are completely accurate. With the help of those powerful software, the
efficiency of CBM well logging data processing and interpretation will be extremely improved.

4. PDF to HTML Conversion: Having a Usable Web Document

Publisher: Muhammad Afzal Bhatti, Adeel Ahmad
Published in: Digital Information Management, 2006 1st International Conference on IEEE Xplore January 2007

Abstract:
In most of the digital libraries and Websites the artifacts are available in portable document format (PDF). Most of the users read online these PDF documents in their daily activities and face problems while reading multi column PDF documents. This paper describes the design of a PDF to HTML converter to ease the online reading activity by providing a usable Web document. PDF to HTML converter, is designed to enhance usability of PDF documents. Our converter presents the document content in a more usable way by following Web usability guidelines. It reduces time and effort required to read PDF documents online.

5. A Highly Accurate PDF-To-Text Conversion System for Academic Papers Using Natural Language Processing Approach

Publisher: Muhammad Afzal Bhatti, Adeel Ahmad

Published in: Digital Information Management, 2006 1st International Conference on IEEE Xplore January 2007

Abstract:
Extracting text out of PDF documents is never an easy task when a higher degree of accuracy and consistency are the two main criteria to be attained. Although, there exist a considerable number of such systems; however, most of them are falling short of offering desirable performance especially when academic literature is the concern. Researches, those involved heavily in text mining and project analyzing, need an accurate and consistent supporting tool for PDF-To-Text (PTT) conversion. Therefore, in this paper, we propose a Natural Language Processing based PDF-to-text (NLPDF) conversion system, which comprises of two major steps, namely (i) reads contents from the PDF and (ii) reconstruct the text.

Chapter 3

Real World Application

3.1 Real world examples of file type conversion applications:
There are lots of real world examples exists, but we cover only few here of several types.

Online Software:

- **https://smallpdf.com/**
  Advantages:
  - It can convert documents file by using server-side language.
  - It is small size application.

  Disadvantage:
  - Privacy Issue (it stores user data on server)
  - Cannot work offline
  - Too much advertisements

- **https://www.coolutils.com/**
  Advantages:
  - It can convert documents, image file by using server-side language.
  - It is small size application.

  Disadvantage:
  - Privacy Issue (it stores user data on server)
  - Cannot work offline
  - Too much advertisements

- **https://ilovepdf.com/**
  Advantages:
  - It can convert documents file by using server-side language.
  - It is small size application.

  Disadvantage:
  - Privacy Issue (it stores user data on server)
  - Cannot work offline
  - Too much advertisements

- **https://online-convert.com/**
  Advantages:
  - It can convert documents, image, video, audio file by using server-side language.
  - It is small size application.

  Disadvantage:
  - Privacy Issue (it stores user data on server)
  - Cannot work offline
  - Too much advertisements

Offline Software:

- **Format Factory**
  Developer: Chen Jun Hao

  Advantages:
  - It can convert, compress all types of file exist.
  - Work offline
  - No privacy issues
  - Free but ad-supported

  Disadvantage:
  - Platform dependent
  - Large size
  - Proper installation guide need by normal user

- **Adobe Acrobat Pro**
  Developer By: Adobe Inc.

  Advantages:
  - It can convert, create, compress document file specially pdf
  - Work offline
  - No privacy issues
  - Costly to purchase

  Disadvantage:
  - Platform dependent
  - Large size
  - Proper installation guide need by normal user

- **Libre Office**
  Developer By: The Document Foundation

  Advantages:
  - It can convert, create document file all types of file exist.
  - Work offline
  - No privacy issues
  - Free open source

  Disadvantage:
  - Platform dependent
  - Large size
  - Proper installation guide need by normal user

Chapter 4

Inferences from Literature

4.1 Principles of Data Conversion:

Before any data conversion is carried out, the user or application programmer should keep a few basics of computing and information theory in mind. These includes:

1. Information can easily be discarded by the computer, but adding new information takes effort.
2. The computer can add information only in a rule-based fashion.
3. Up-sampling the data or converting to a more feature-rich format does not add information; it merely makes room for that addition, which usually a human must do (Automatic restoration of information that was lost through a lossy compression process would probably require important advances in artificial intelligence).
4. Data stored in an electronic format can be quickly modified and analysed.

Because of these realities of computing and information theory, data conversion is often a complex and error-prone process that requires the help of experts.

4.2. Different Types of File Formats:

1. **Open Formats:**
   An **open format** is a file format for storing digital data, defined by a published specification usually maintained by a standards organization, and which can be used and implemented by anyone. For example, an open format can be implemented by both proprietary and free and open-source software, using the typical software licenses used by each. In contrast to open formats, closed formats are considered trade secrets. Open formats are also called **free file formats** if they are not encumbered by any copyrights, patents, trademarks or other restrictions (for example, if they are in the public domain) so that anyone may use them at no monetary cost for any desired purpose.

   Open formats examples:

   - PNG — a raster image format standardized by ISO/IEC
   - HTML — Hypertext Mark-up Language (HTML) is the main mark-up language for creating web pages and other information that can be displayed in a web browser.
   - gzip — for compression
   - JPEG 2000 – an image format standardized by ISO/IEC

2. **Proprietary Formats:**

A **proprietary format** is a file format of a company, organization, or individual that contains data that is ordered and stored according to a particular encoding-scheme, designed by the company or organization to be secret, such that the decoding and interpretation of this stored data is easily accomplished only with particular software or hardware that the company itself has developed. The specification of the data encoding format is not released, or underlies non-disclosure agreements. A proprietary format can also be a file format whose encoding is in fact published, but is restricted through licenses such that only the company itself or licensees may use it. In contrast, an open format is a file format that is published and free to be used by everybody.
Proprietary formats examples:

- PSD – (documented) Adobe Photoshop's native image format
- RAR – (partially documented) archive and compression file format owned by Alexander L. Roshal
- WMA – a closed format, owned by Microsoft

4.3 Image and Images File Format

What is Image File Format?

Image file formats are standardized means of organizing and storing digital images. An image file format may store data in an uncompressed format, a compressed format (which may be lossless or lossy), or a vector format. Image files are composed of digital data in one of these formats so that the data can be rasterized for use on a computer display or printer. Rasterization converts the image data into a grid of pixels. Each pixel has a number of bits to designate its color (and in some formats, its transparency). Rasterizing an image file for a specific device considers the number of bits per pixel (the color depth) that the device is designed to handle.

1. PNG
Basic Definition: PNG or Portable Network Graphics files are a lossless image format originally designed to improve upon and replace the gif format. PNG files are able to handle up to 16 million colours, unlike the 256 colours supported by GIF. A raster image format standardized by ISO/IEC.
Compression: Lossless - compression without loss of quality
Best For: Web Images
Special Attributes: Save Transparency
Open format
MIME media type `image/png`
Conversion: lossy
encoding: UTF-8

2. JPG
Basic Definition: JPEG, which stands for Joint Photographic Experts Groups is a “lossy” format meaning that the image is compressed to make a smaller file. The compression does create a loss in quality but this loss is generally not noticeable. JPEG files are very common on the Internet and JPEG is a popular format for digital cameras - making it ideal for web use and non-professional prints. An image format standardized by ISO/IEC

Compression: Lossy - some file information is compressed or lost
Best For: Web Images, Non-Professional Printing, E-Mail, PowerPoint
Special Attributes: Can choose amount of compression when saving in image editing programs like Adobe Photoshop or GIMP.
Open format
MIME media type `image/jpeg`
encoding: UTF-8

3. GIF
Basic Definition: GIF or Graphics Interchange Format files are widely used for web graphics, because they are limited to only 256 colours, can allow for transparency, and can be animated. GIF files are typically small is size and are very portable.

Compression: Lossless - compression without loss of quality
Best For: Web Images
Special Attributes: Can be Animated, Can Save Transparency
Formerly proprietary now Open format
MIME media type `image/gif`
encoding: UTF-8

4. WEBP
Basic Definition: A WEBP file is an image saved in the WebP (pronounced "Weppy") raster image format developed by Google for web graphics. The WebP format reduces file size more than standard JPEG compression while maintaining similar or better image quality.

Compression: supports both lossy and lossless compression
Best For: Web Images (not supported by safari, Internet Explorer)
Special Attributes: includes an alpha channel for transparency. Several graphics editors can open and save WebP files, such as Adobe Photoshop, Gimp, Image Magick, and IrfanView.
License: developed by [Google](https://en.wikipedia.org/wiki/Google), based on technology acquired with the purchase of [On2 Technologies]
MIME media type `image/webp`
encoding: UTF-8

5. SVG
Basic Definition: An SVG file is a graphics file that uses a two-dimensional [vector graphic] format created by the World Wide Web Consortium (W3C). It describes images using a text format that is based on [XML](https://techterms.com/definition/xml). SVG files are developed as a standard format for displaying vector graphics on the web.

Compression: supports both lossy and lossless compression
Best For: Web pages
Special Attributes: SVG images can be created and exported from Adobe Creative Suite programs, such as Illustrator and Go Live.
License: open standard by w3c Open Format
MIME media type `image/svg+xml`
encoding: UTF-8, XML

4.4 Document and Document File Format

What is document file format?

A document file format is a text or binary file format for storing documents on a storage media, especially for use by computers. There currently exist a multitude of incompatible document file formats. A rough consensus has been established that XML is to be the technical basis for future document file formats, although PDF is likely to remain the format of choice for fixed-layout documents. Examples of XML-based open standards are Doc Book, XHTML, and, more recently, the ISO/IEC standards OpenDocument (ISO 26300:2006) and Office Open XML (ISO 29500:2008). HTML is the most used and open international standard and it is also used as document file format. It has also become ISO/IEC standard (ISO 15445:2000). The default binary file format used by Microsoft Word (.doc) has become widespread *de facto* standard for office documents, but it is a proprietary format and is not always fully supported by other word processors.

1. DOCX
Basic Definition: A DOCX file is a document created by Microsoft Word or another [word processing] (https://techterms.com/definition/wordprocessor) program, such as OpenOffice Writer or Apple Pages. It contains formatted text but may also include images, drawn objects, and other document elements. DOCX files are widely used in home, academic, and business environments for drafting letters, resumes, invitations, newsletters, and other documents.

NOTE: To explore the contents of a DOCX file manually, rename the ".docx" extension to ".zip" and then decompress the resulting file with any zip decompression utility.

Compression: lossless
License to Microsoft
MIME media type `application/vnd.openxmlformats-officedocument.wordprocessingml.document`

Used by Application: Word 2007 or later, LibreOffice

2. PDF
Basic Definition: A PDF file is a multi-platform document created by Adobe Acrobat or another PDF application. The PDF format is commonly used for saving documents and publications in a standard format that can be viewed on multiple platforms. The files may contain text, images, forms, annotations, outlines, and other document-related data. The PDF files also preserve fonts and formatting electronically across multiple platforms and appear the same on the screen as when printed on paper.

Used by Application: Adobe Reader, LibreOffice, Apple Pages, and web browsers, such as Google Chrome and Microsoft Edge.

MIME media type `application/pdf`

Formerly proprietary but now free open source

3. TXT
Basic definition: A TXT file is a standard text document that contains unformatted text. It is recognized by any text editing or word processing program and can also be processed by most other software programs. TXT files are often created and opened using Microsoft Notepad and Apple TextEdit.

Open Format
MIME media type `text/plain`
Encoding: UTF-8/ASCII

Used by Application: Web browser, text editors

4. HTML
Basic definition: An HTML file is a web page coded in HTML that can be displayed in a web browser. It is used to format text, tables, images, and other content that is displayed on a web page. HTML files are widely used on the web as most pages within [static websites](https://techterms.com/definition/staticwebsite) have an ".html" extension. HTML source code is parsed by a web browser and is typically not seen by the user.

Open Format
MIME media type `text/html`
Encoding: UTF-8
Used by Application: Web browser, text editors

5. MD
Basic definition: An MD file is a text file created using one of several possible dialects of the Markdown language. It is saved in plain text format but includes in-line text symbols that define how to format the text (e.g., bold, indentations, headers, table formatting). MD files are designed for authoring plain text documentation that can be easily converted to HTML. Projects created with GitHub, a popular on-line version control system, often use a file named README.md, which contains documentation for the project.
John Gruber, the developer of Markdown, has created a Perl program for converting Markdown to HTML. The program, called "Markdown," is available at the Daring Fireball website.

Open Format
MIME media type `text/markdown`
Used by Application: Github, Software Documentation, text editors

4.5 Audio and Audio File Format

What is audio file format?

An audio file format is a file format for storing digital audio data on a computer system. The bit layout of the audio data (excluding metadata) is called the audio coding format and can be uncompressed, or compressed to reduce the file size, often using lossy compression. The data can be a raw bitstream in an audio coding format, but it is usually embedded in a container format or an audio data format with defined storage layer.

1. MP3
Basic definition: An MP3 file is an audio file saved in a compressed audio format developed by the Moving Picture Experts Group (MPEG) that uses "Layer 3" audio compression. It is commonly used to store music and audiobooks with near-CD quality sound (stereo, 16-bit) and roughly 1/10 the size of a .WAV or .AIF file. The quality of an MP3 file depends largely on the bit rate used for compression. Common bit rates are 128, 160, 192, and 256 kbps. Higher bit rates result in higher quality files that also require more disk space.

Device support: Apple iPod and Sony Walkman devices, various portable music devices
Open Format
Compression: lossy
MIME media type `audio/mpeg`
Encoding: MPEG-1

2. WAV
Basic definition: A WAV file is an audio file saved in the WAVE format, which is a standard digital audio file format utilized for storing waveform data. WAV files may contain audio recordings with different sampling rates and bit rates but are often saved in a 44.1 KHz, 16-bit, stereo format, which is the standard format used for CD audio. The WAVE format is based on the Resource Interchange File Format (RIFF), which is a file container format primarily used for saving video and sound. Microsoft and IBM jointly developed the WAVE format in the early-1990s adopting WAV files as the primary option for saving raw, uncompressed audio in Microsoft Windows.

Encoding: RIFF
Proprietary format from Microsoft and IBM
MIME media type `audio/wav`
Compression: lossless

4.6 Compression and Compressed File Formats

What is compressed file formats?

A compressed file is any file that contains one or more files or directory that is smaller than their original file size. These files make downloading faster easier and allow more data to be stored on a removable media. Common compressed file extensions are .ZIP, .RAR, .ARJ, TAR.GZ, and .TGZ. Below is a larger listing of the different compressed file extensions you are likely to come across when working on a computer.

1. ZIP
Basic definition: The Zip format was created in 1989 by Phil Katz and was used in the PKZIP utility developed by PKWARE, Inc. The format grew in popularity and is now supported by most file compression/decompression programs. It stores files separately from each other allowing the files to be compressed using different methods and extracted without compressing or decompressing the entire archive.

Programs to open ZIP: WinRAR, 7zip, Microsoft file explorer, etc

MIME media type `application/zip`
Open Format
compression: lossless

2. TAR.GZ
Basic definition: A TAR.GZ file is a TAR archive compressed with the standard GNU zip (gzip) compression algorithm. It contains one or more compressed files and is commonly used on Unix operating systems to package files, programs, and installers.
NOTE: TAR.GZ files must first be decompressed and then expanded using a TAR utility. They include both .TAR and .GZ file types.

Programs to open tar.gz: WinRAR, 7zip, Ark, etc.
Open Format
MIME media type `application/gzip`

3. RAR
Basic definition: A RAR file is an archive that contains one or more files compressed with RAR compression. It uses a higher compression ratio than typical ZIP compression and incorporates a proprietary compression algorithm. The RAR compressor can also create spanned, or multi-volume, archives, which are split across several compressed files. These files typically have file extensions from ".R00" to ".R99," or they may all have the standard ".RAR" extension.
NOTE: The name "RAR" stands for "Roshal Archive," which was named after Eugene Roshal, the RAR compression algorithm developer.

Programs to open rar: WinRAR, 7zip, unrar, etc.
Proprietary to win.rar GmbH
MIME media type `application/vnd.rar`

Chapter 5

Requirements

5.1 Hardware Requirements

- Computer System – Laptop, desktop, mobile device
- Web browser – Chrome, Firefox, Edge

5.2 Software Requirements

- Web Technologies – NodeJS, ReactJS, Cache Storage
- Other – Git, Github, NPM, VSCODE Editor, etc

Bit About Software Technologies Used:

- **Node.js** was initially built for Google Chrome, and later open-sourced by Google in 2008. It is built on Chrome's V8 JavaScript engine. It's designed to build scalable network applications, and can execute JavaScript code outside of a browser.Node.js works without an enclosing HTML page, instead using its own module system based on CommonJS, to put together multiple JavaScript files.

- **React.js** was originally created by a software engineer at Facebook, and was later open-sourced. It is maintained by Facebook, as well as a community of development companies and individual developers.
  The React library can be used for creating views rendered in HTML. React views are declarative. This means that developers don't have to worry about managing the effects of changes in the view's state (the object that determines how components behave) or changes in the data.

- **Git** is a distributed version-control system for tracking changes in source code during software development. It is designed for coordinating work among programmers, but it can be used to track changes in any set of files. Its goals include speed, data integrity, and support for distributed, non-linear workflows.
  Git was created by Linus Torvalds in 2005 for development of the Linux kernel, with other kernel developers contributing to its initial development. Its current maintainer since 2005 is Junio Hamano. As with most other distributed version-control systems, and unlike most client–server systems, every Git directory on every computer is a full-fledged repository with complete history and full version-tracking abilities, independent of network access or a central server. Git is free and open-source software distributed under the terms of the GNU General Public License version 2.

- **GitHub, Inc.** is a US-based global company that provides hosting for software development version control using Git. It is a subsidiary of Microsoft, which acquired the company in 2018 for US$7.5 billion. It offers the distributed version control and source code management (SCM) functionality of Git, plus its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project.

GitHub offers plans free of charge, and professional and enterprise accounts. Free GitHub accounts are commonly used to host open source projects. As of January 2019, GitHub offers unlimited private repositories to all plans, including free accounts. As of January 2020, GitHub reports having over 40 million users and more than 100 million repositories (including at least 28 million public repositories), making it the largest host of source code in the world.

- **CacheStorage** is a storage mechanism in browsers for storing and retrieving network requests and response. It stores a pair of Request and Response objects. The Request as the key and Response as the value. The Cache Storage API opens up a whole new range of possibilities, by giving developers complete control over the contents of the cache. Instead of relying on a combination of HTTP headers and the browser's built-in heuristics, the Cache Storage API exposes a code-driven approach to caching. The Cache Storage API is particularly useful when called from your service worker's JavaScript.
- **NPM** (originally short for Node Package Manager) is a package manager for the JavaScript programming language. It is the default package manager for the JavaScript runtime environment Node.js. It consists of a command line client, also called npm, and an online database of public and paid-for private packages, called the npm registry. The registry is accessed via the client, and the available packages can be browsed and searched via the npm website. The package manager and the registry are managed by npm, Inc.
- **Google Chrome** is a cross-platform web browser developed by Google. It was first released in 2008 for Microsoft Windows, and was later ported to Linux, macOS, iOS, and Android. The browser is also the main component of Chrome OS, where it serves as the platform for web apps. Most of Chrome's source code comes from Google's open-source Chromium project, but Chrome is licensed as proprietary freeware. Web Kit was the original rendering engine, but Google eventually forked it to create the Blink engine; all Chrome variants except iOS now use Blink. As of July 2019, Stat Counter estimates that Chrome has a 71% worldwide browser market share on traditional PCs and 63% across all platforms. Because of this success, Google has expanded the "Chrome" brand name to other products: Chrome OS, Chromecast, Chromebook, Chromebit, Chromebox, and Chromebase.
- **Visual Studio Code**

  It is a source-code editor developed by Microsoft for Windows, Linux and macOS. It includes support for debugging, embedded Git control and GitHub, syntax highlighting, intelligent code completion, snippets, and code refactoring. It is highly customizable, allowing users to change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality. The source code is free and open source and released under the permissive MIT License. The compiled binaries are freeware and free for private or commercial use. Visual Studio Code is based on Electron, a framework which is used to develop Node.js applications for the desktop running on the Blink layout engine. Although it uses the Electron framework, the software does not use Atom and instead employs the same editor component (codenamed "Monaco") used in Azure DevOps (formerly called Visual Studio Online and Visual Studio Team Services). In the Stack Overflow 2019 Developer Survey, Visual Studio Code was ranked the most popular developer environment tool, with 50.7% of 87,317 respondents claiming to use it.

Chapter 6

Proposed solution

6.1 Proposed work

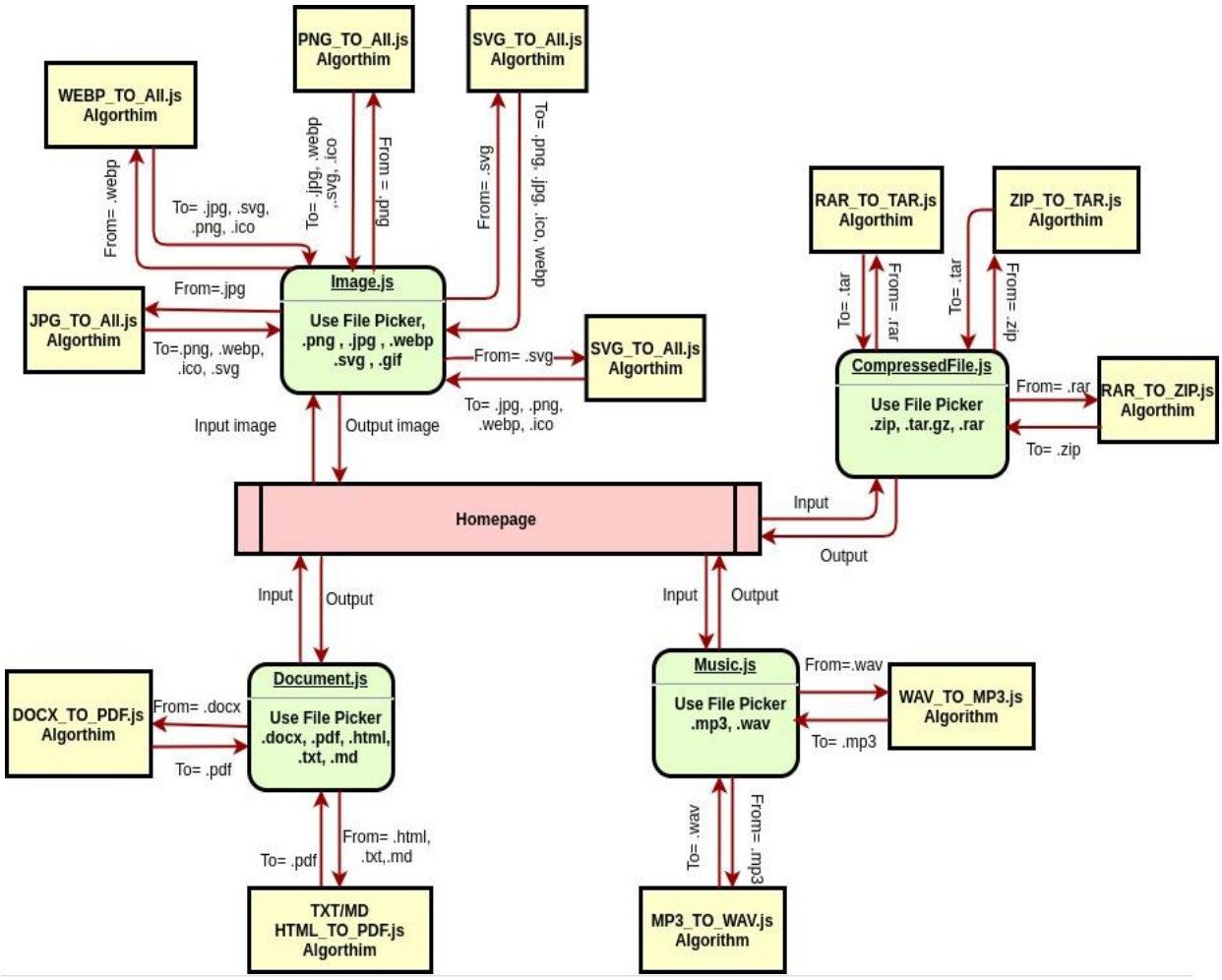
In our proposed model we have worked with the following file algorithms:

- Image File Conversion: JPG, PNG, SVG, GIF, WEBP
- Document File Conversion: PDF, DOCX, TXT, HTML, MD
- Audio File Conversion: MP3, WAV
- Compressed File Conversion: ZIP, RAR, TAR

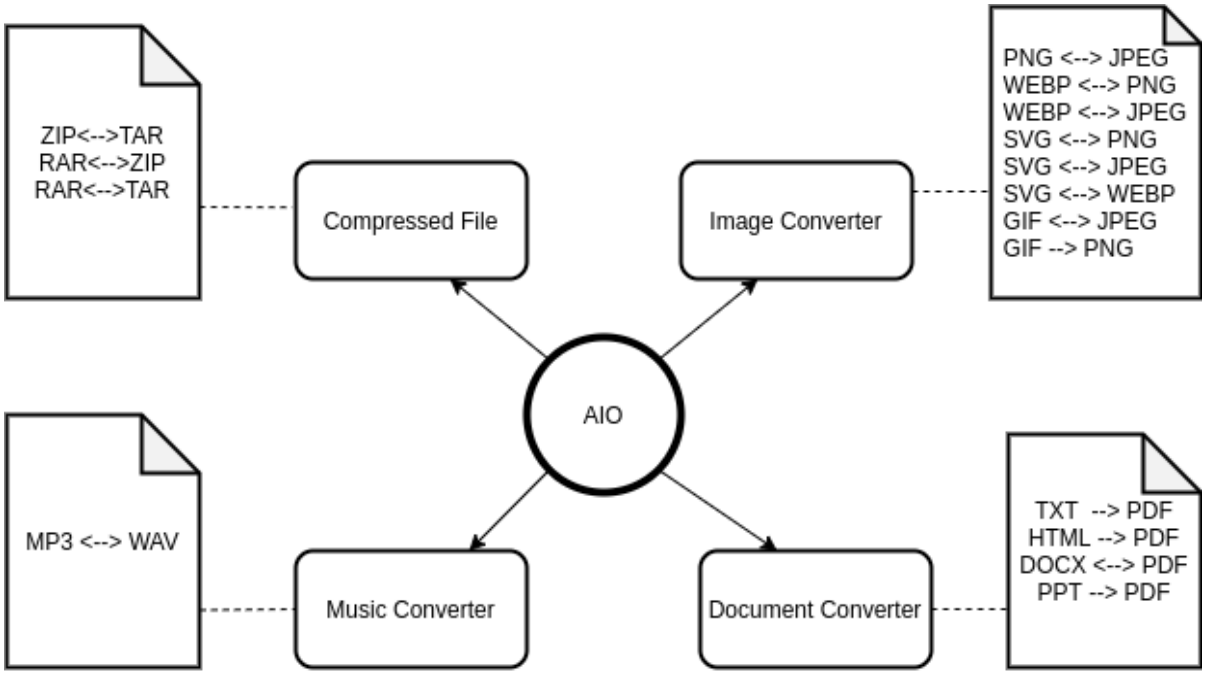
Some open source algorithm, some self-created algorithm, some modified and combinational algorithm are used in this project.

6.2 Design

Data Flow Diagram:



Component Diagram:



Component Diagram

6.3 Algorithms

Images Algorithms:

PNG -> JPG/WEBP
JPG -> PNG/WEBP
GIF -> PNG/JPG/WEBP
WEBP -> JPG/PNG
SVG -> JPG/PNG/WEBP

```
* 1-> Converting image to base64.
* 2-> pass base64 to function.
* 3-> create canvas with that base64 data uri.
* 4-> set with window.innerHieght and width, so that according to image size.
* 5-> read the canvas image and change image signature to respective file type
* 6-> convert canvas to image by toDataUrl() function in js.
```

Document Algorithms:

TXT -> PDF

```
* 1-> Read txt file as plain text.
* 2-> According to length of text content, decide the number of pages.
* 3-> Then by using Open Source library jsPDF() passing arguments as text data.
* 4-> jsPDF() return a BLOB data, save it as pdf.
```

HTML -> PDF

```
* 1-> Read html file as plain text.
* 2-> According to length of html content, decide the number of pages.
* 3-> Then by using Open Source library jsPDF() passing arguments as text data.
* 4-> jsPDF() return a BLOB data, save it as pdf.
```

MD -> PDF

```
* 1-> Read html file as plain text.
* 2-> According to length of html content, decide the number of pages.
* 3-> Then by using Open Source library jsPDF() passing arguments as text data.
* 4-> jsPDF() return a BLOB data, save it as pdf.
```

Audio Algorithms:

MP3 -> WAV

Compressed File Algorithms:

ZIP -> Tar
Tar -> ZIP

6.4 Source Code and Project Setup

Complete Project Code: https://github.com/MasterKN48/aio/
Live Project Link: https://masterkn48.github.io/aio/

Project Setup:

- Git clone the project from Github project link.

- In the project directory, you can run:
  `npm start`
  Runs the app in the development mode.

- Open (http://localhost:3000) to view it in the browser.
  The page will reload if you make edits.
  You will also see any lint errors in the console.

- `npm test`
  Launches the test runner in the interactive watch mode.
  See the section about [running tests] (https://facebook.github.io/create-react-app/docs/running-tests) for more information.

- `npm build`
  Builds the app for production to the `build` folder.
  It correctly bundles React in production mode and optimizes the build for the best performance.

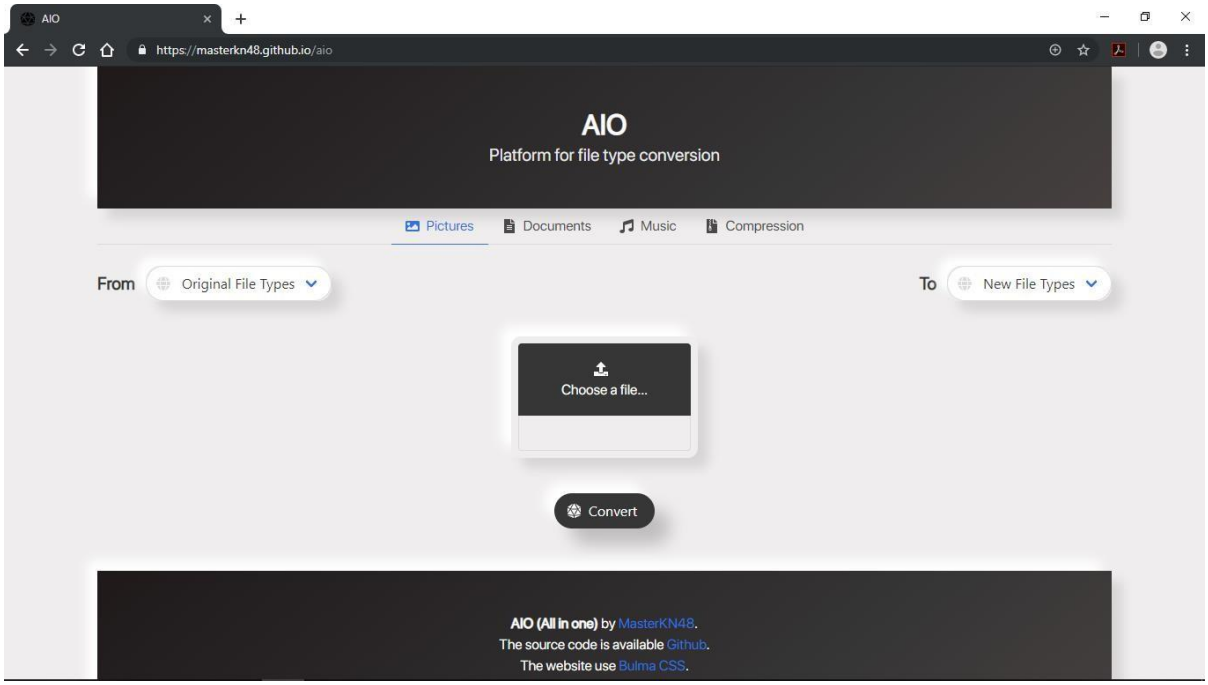
  The build is minified and the filenames include the hashes
  Your app is ready to be deployed!

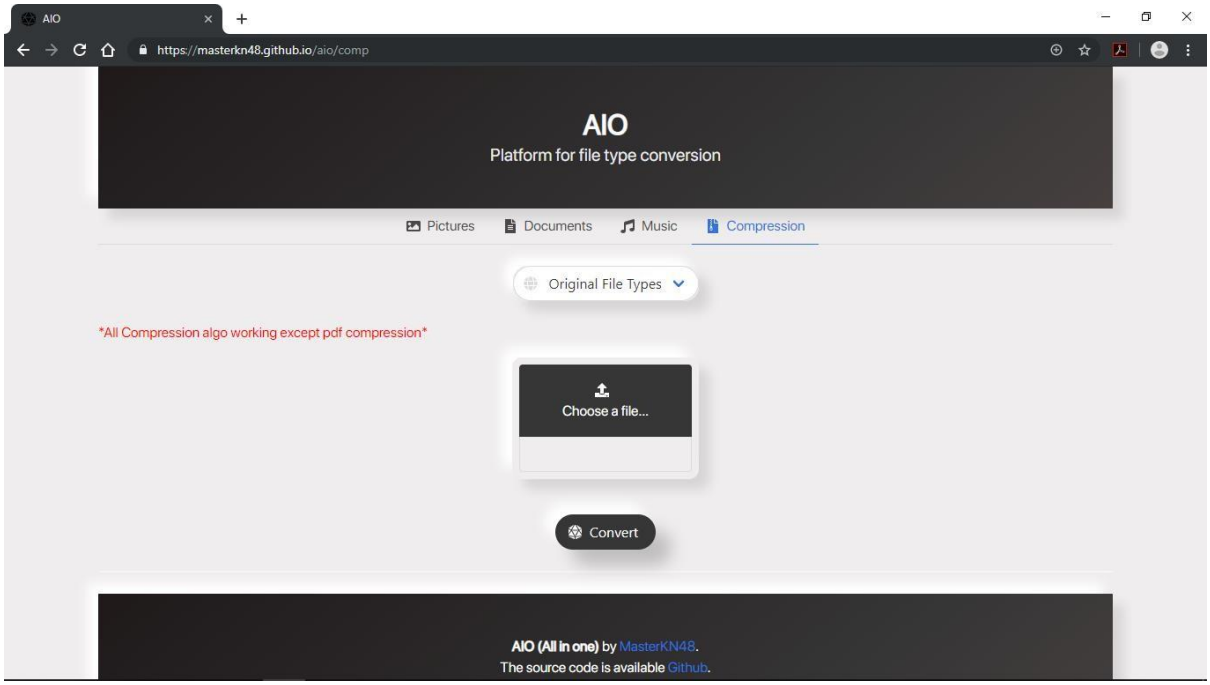
- `npm eject`
  Note: this is a one-way operation. Once you `eject`, you can’t go back!
  If you aren’t satisfied with the build tool and configuration choices, you can `eject` at any time. This command will remove the single build dependency from your project.

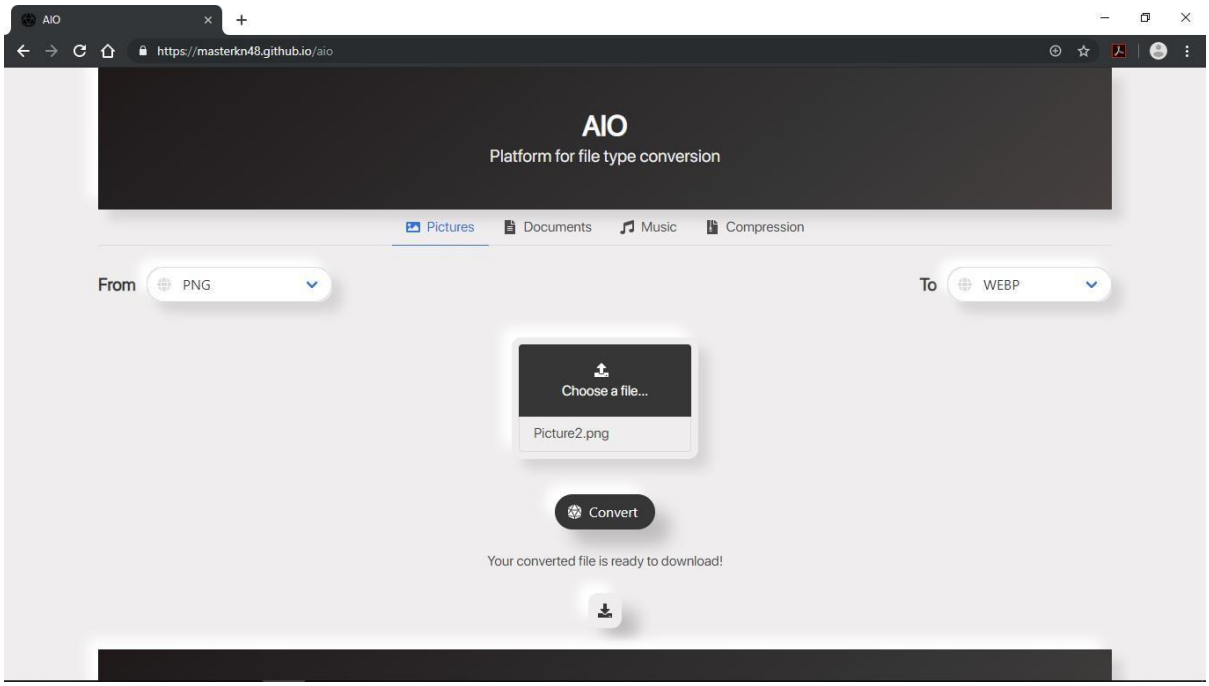
- Instead, it will copy all the configuration files and the transitive dependencies (Webpack, Babel, ESLint, etc.) right into your project so you have full control over them. All of the commands except `eject` will still work, but they will point to the copied scripts so you can tweak them. At this point you’re on your own.

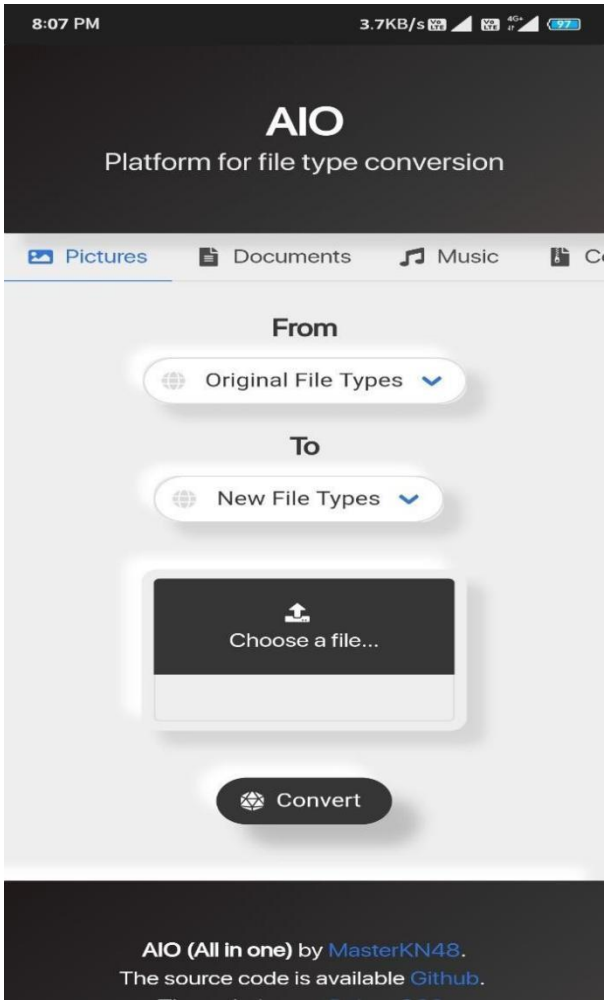
- You don’t have to ever use `eject`. The curated feature set is suitable for small and middle deployments, and you shouldn’t feel obligated to use this feature. However, we understand that this tool wouldn’t be useful if you couldn’t customize it when you are ready for it.

6.5 Sample/Screenshot



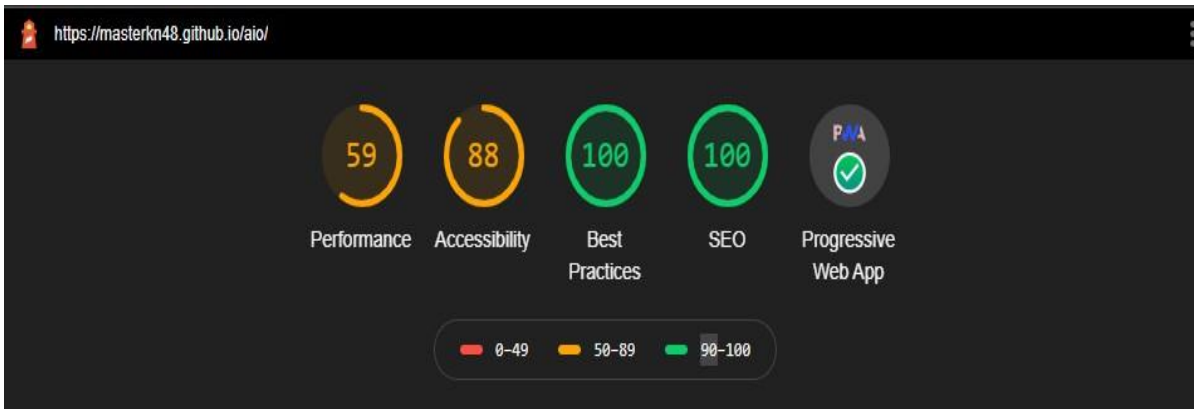



AIO
https://masterkn48.github.io/aio
AIO
Platform for file type conversion
Pictures
Documents
Music
Compression
From
PNG
To
WEBP
Choose a file...
Picture2.png
Convert
Your converted file is ready to download!


8:07 PM
3.7KB/s
AIO
Platform for file type conversion
Pictures
Documents
Music
From
Original File Types
To
New File Types
Choose a file...
Convert
AIO (All in one) by MasterKN48.
The source code is available Github.

6.6 Performance Metrics

This test is done in Google Chrome Audits Tool



Chapter 7

Proposed Time bound progress

7.1 Progress Sheet

Milestones	Date Assigned	Date Completed	Task
1. Study over file conversion	5/01/2020	Till the project end	Study about various aspects of File formats and conversion.
2. Designing the application	10/01/2020	15/01/2020	Task is to create design of application like wireframe, mockups, understanding design challenges.
3. Modules Generation	16/01/2020	20/01/2020	Deciding number of modules 1-> Image Convertor, 2-> Document Convertor, 3-> Compressed File Convertor, 4-> Audio Convertor
4. Creating Image Convertor	21/01/2020	02/02/2020	Searching about images conversion algorithms, and created all image convertor
5. Create Document Convertor	03/02/2020	10/02/2020	Creating document handling modules.
6. Document Modules	11/02/2020	15/02/2020	Study about documents file format. Creating algorithm design for txt to pdf.
7. Document Modules	16/02/2020	20/02/2020	Study about open formats, proprietary formats, worked on html to pdf and pdf to html.
8. Creating File Compression Modules and documents modules	21/02/2020	30/02/2020	Searching about compression, conversion algorithms, and creating algorithms for file compression

Chapter 8

Conclusion and Future Work

8.1 Conclusions

The Purpose and objective of our project is to provide security to user who doesn't want to share crucial data to anyone else but need a platform to convert their file without any fear of data privacy.
This project will provide a platform which can available anywhere and be use at any time. This project will provide small size very fast file convertor. Creating a secure and free platform for the file format conversion which supports document, audio, image conversion along with the added feature for the file compression. This platform use client-side javascript for file conversion it does not going to store users file on anywhere on server. Server will use only to host web application for anywhere access. All the Computation will be done at client side by user machine only.

8.2 Future Work

In this research a file conversion system has been designed for converting a file from one format to another. Our project target a lot of data formats in various categories of File formats, but there is a need to do more comprehensive observations and activities. Here are some of them:

- In our proposed model we are able to convert a text files of format TXT ,MD, DOCX, and PDF. We would like to target more Formats in future like CSV, EPUB ,DOTX, etc

- In our proposed model we are able to convert a image files of format PNG ,SVG and JPEG. We would like to target more Formats like AI , PSD,CAD

- we would like to provide a broader support for more audio, video and zip file formats.

• System performance should be evaluated in a scalable environment in order to measure how responsive it is in case of large amount of service requests. This will also show how resistant the system is against denial of service attacks.

References

[1] PC World (23 December 2003). "Windows Tips: For Security Reasons, It Pays To Know Your File Extensions". Archived from the original on 23 April 2008. Retrieved 20 June 2008.

[2] R. Mohemad, A. R. H. Z. A. Othman, N. M. M. Noor. Automatic document structure analysis of structured PDF files. International Journal of New Computer Architectures and their Applications (IJNCAA), 1(2)(2011), 404-411.

[3] R. A. Amsler. The structure of the merriam-webster pocket dictionary. (1980). Doctoral Dissertation.

[4] Printernational. ISO Paper sizes. (2017). Available: http://www.printernational.org/iso-paper-sizes.php

[5] Cambridge Computer Laboratory. (2017). Available: https://www.cl.cam.ac.uk/~mgk25/iso-paper-ps.txt

[6] X. Xueya, Z. Yanmei. The research and application of the creation PDF document based on the iTextSharp. Proceedings of the IEEE Symposium on Robotics and Applications (ISRA),Kuala Lumpur,Malaysia, (2012).

[7] S. Bird. NLTK: the natural language toolkit. Proceedings of the COLING/ACL on Interactive presentation, Sydney, Australia, (2006).

[8] Kamble, A. Jaiswal, N. Dekate, S. Haridas, K. Pendke. Integrated System for Reading Multiple Files. International Journal of Computer Science and Mobile Computing, 3(2)(2014),45-52.

[9] R. Milton. Extracting Data from PDFs: Clean Air in Schools.(2011). Available: http://genesis.blogs.casa.ucl.ac.uk/2011/07/18/extracting-data-from-pdfs-clean-air-in-schools/.

[10] V. Gaikar. Zamzar: Convert Your Files to Any Format by Email.Tricks Machine, (2012).Available:http://www.tricksmachine.com/2012/03/zamzar-convert-yourfiles-by-mail.html/

[11] Z. Ahmed, T. Dandekar. MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format. F1000Research, 4, (2015).

[12] Mediafox Marketing s.r.o. PDF to Text. (2017). Available: http://pdftotext.com/

[13] Online2PDF.com. Online2PDF.com 8.0. (2017). Available:https://online2pdf.com/

[14] Online-convert.com. Free online file converter. (2017). Available: http://online-convert.com/

[15] G. A. Miller. WordNet: A Lexical Database for English. Communications of the ACM, 38(11)(1995), 39-41

[16] C. Fellbaum. WordNet: An Electronic Lexical Database. (1998). Cambridge, MA, MIT Press