Challenges in Data Conversion Process

Ansh Sachdeva, Nitin Kumar, Kunal Wadhwan

B.Tech CSE, JIMS Engineering Management Technical Campus (JEMTEC), GGSIP University
New Delhi, India

anshsachdeva2013@gmail.com
nitinkumar@ieee.org
kunalwadhwan46@gmail.com

Abstract: **Computer files are crucial with respect to Data storage and interpretation. Throughout a computer environment, data is encoded in a variety of ways. Thus, data must be converted in some form before it can be used by a different computer, operating system or program. This document gives a general overview regarding various Challenges that arise during the process of conversion from one format to another.**

Keywords: Data, File ,File Format, Encoding , Data Conversion, Lossless, Lossy Conversion.

INTRODUCTION

As per NIST Computer file is a computer resource for recording data discretely in a computer storage device. There are different types of computer files, designed for different purposes. A file may be designed to store a picture, a written message, a video, a computer program, or a wide variety of other kinds of data. Some types of files can store several types of information at once [1].

A file format is a standard way of defining how the information is encoded for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free and may be either unpublished or openly specified. A file format is standard can be used to provide various file related information:

- **Compression Technique Used:** The data in a file can be either compressed (lossy or lossless) or uncompressed [2] .
- **Encoding used**: The content of file could also be present as plain binary data or encoded/encrypted in some standard encoding (like ascii or UTF-8 or html).
- **Data Uniformity:** The file could also be holding a specific type of data(for e.g. PNG files only store binary encoded data; a txt file only stores ascii encoded text) or holding multiple types of data(for e.g. an .mp4 file could hold audio data, video data and also metadata*)
- **File Structure**: The content of file might be following a particular file structure
- **File Attributes**: Multiple file attributes could be attached to a file
- **Permission Levels**: The file could be locked to provide read/write/modify access to certain user groups/all users/ users with a key

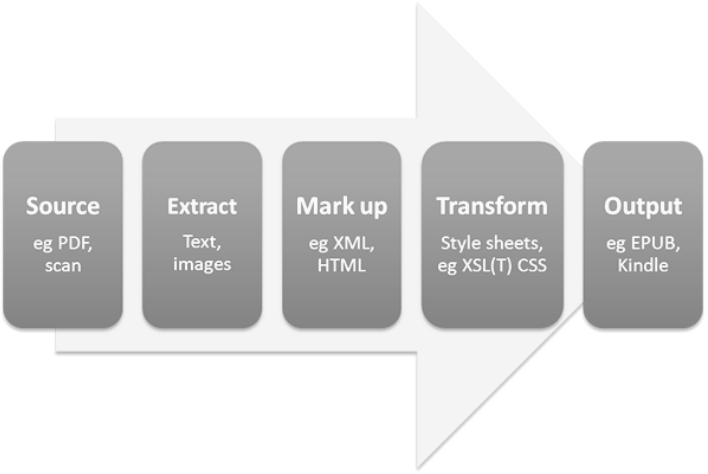

Figure 1 :A basic pdf to ePUB conversion process

Principles of Data Conversion:

Before any data conversion is carried out, the user should know the basics rules 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 [3]. 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.

For example, a colourful image can easily be converted to grayscale, while the opposite conversion is a nearly impossible process without involving

artificial intelligence[4]. Grey scaling an image is rule-based, whereas the addition of colour information to a grayscale image cannot be done programmatically [5] . This is because s only a human knows which colours are needed for each section of the picture–there are no rules that can be used to automate that process [6] .

Even Upscaling programmes like Converting a 24-bit PNG to a 48-bit one would not add information to it, it would only pads existing RGB pixel values with zeroes, so that a pixel with a value of DD A4 23, for example, becomes DD00 A400 2300 [7].

Converting an image or audio file in a lossy format (like JPEG or Vorbis) to a lossless (like PNG or FLAC) or uncompressed (like BMP or WAV) format only wastes space, since the same image with its loss of original information (the artefacts of lossy compression) becomes the target [8]. A JPEG image can never be restored to the quality of the original image from which it was made, no matter how much the user tries the "JPEG Artefact Removal" feature of their image manipulation program [9].

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.

Challenges in Data conversion

There are many ways in which data is converted within the computer environment:

1. Complexity in Procedure of Conversion

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 coloured image to grayscale) [10-13]. It can also be as complex as conversion between different image formats, which requires deeper understanding of compression and encoding algorithms.

2. Direct and Pivotal conversion

Data conversion can occur directly from one format to another, but many applications that convert between multiple formats use an intermediate representation by way of which any source format is converted to its target. This is called pivotal conversion. for e.g., for a .DOC to .PDF conversion a .DOC file is first converted to a proprietary xml format, which is then compressed and encoded to a .PDF file[14] . An audio converter that converts from FLAC to AAC decodes the source file to raw PCM data in memory first, and then performs the lossy AAC compression on that memory image to produce the target file.

3. Multi-pass conversion

A conversion could happen in a single stage or it may involve a complex process of going through intermediary stages, which would result int longer time taken for conversion [15] .

4. Lost and inexact data conversion.

The objective of data conversion is to convert data from one format to another maintaining as much of the embedded information and data as possible but oftentimes this goal is not achieved due to various factors.
One reason could be the lack of support of same features and data structures between the source and target format. If the source format has features that are not yet supported by the target format, the data will be lost [16] . For e.g., converting a word file (.DOC) to a text file (.TXT) will result in a lot of data being lost, simply because text files do not support as many features as word files.

Another cause of data loss could be inexactitude, which means the incomplete output received due to conversion between formats that are conceptually different. For e.g., the word processors in Microsoft office use WYSIWYG paradigm [1] in their document creation applications, while many other document editors like HTML, PDF and MathML editors use structural descriptive paradigms found in SGML and XML. That is why a direct conversion from .PDF to editable word processor file would result in an inexact conversion (a word “HELLO” with spacing of 1em is converted to “H E L L O” on a word file)

5. Proprietary codecs involved

A proprietary format, in contrast to open format, is a file format for storing digital data according a particular encoding scheme designed to be only decoded and/or interpreted by the software of the company behind it. Proprietary formats are typically controlled by a company or organization for its own benefits, and the restriction of its use by others is ensured through patents or as trade secrets. It is thus intended to give the license holder exclusive control of the technology to the (current or future) exclusion of others [16] .

This is a challenge to data conversion, as a successful data conversion would require thorough knowledge of the workings of both source and target formats. In the case of proprietary formats, reverse engineering is carried out for conversion. Reverse engineering can achieve close approximation of the original specifications, but errors and missing features can still be present in the final result.

6. Security Risks involved.

File conversion is a resource intensive process. Converting a raw Blu-ray video to a compressed .MP4 format video can take long hours even on a high spec machine. That is why online solutions usually deploy these resource intensive processes on cloud servers to reduce the conversion time. However, this poses a great security risk because a copy of user's file is now handled by multiple parties and there is no transparency involved in the conversion process.

CONCLUSION

In this paper, the various challenges that could occur in the process of converting a data file from one format to another is discussed. Converting a file from one format to another can be very simple if the standards involved are open and actively maintained. But for proprietary formats, reverse engineering is oftentimes the final solution. Sometimes a conversion also ivolevs a lot of intermediate steps and passes. The conversion will always result in a failure or inexact output, if the target format is incompatable with the source format or deos not support all of its features. finally, the security threats that are the result of using internet enabled services for conversion of a file from one format to another are also discussed.

ACKNOWLEDGMENT

The authors would like to thank Dr. Vikas Chowdhary, HOD,CSE Dept. and Mr. Pankaj Singh Yadav, Sr Professor , JIMS, GGSIP University, for their useful discussions and suggestions during the preparation of this technical paper.

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