LONG-TERM PRESERVATION OF PDF FILES IN INSTITUTIONAL REPOSITORIES IN JAPAN

Teru Agata
Asia University
Japan
agata@asia-u.ac.jp
0000-0002-1593-9323

Yosuke Miyata
Teikyo University
Japan
m@miyay.org
0000-0002-5239-5396

Atsushi Ikeuchi
University of Tsukuba
Japan
atsushi@slis.tsukuba.ac.jp
0000-0002-2601-3892

Abstract – In the open access environment, many textual resources have become available in the PDF format on the Web. This research aims to survey PDF files in Japanese institutional repositories (IRs) to address the problems encountered during their long-term preservation. With that aim, 1.5 million PDF files collected from Japanese IRs were analyzed with regard to file format, encryption, and metadata. Most PDF files did not conform to PDF/A. A total of 30.5% of PDFs were encrypted and many PDFs did not have embedded metadata. These results imply that PDF files in Japanese IRs have several serious problems for their long-term preservation.

Keywords – PDF, Metadata, Institutional Repositories, Long-term preservation

Conference Topics – Building Capacity, Capability and Community

I. INTRODUCTION

In the open access environment, many textual resources have become available in PDF format on the Web. It is desirable for science to sustain an open access environment. If IRs were impermanent, it becomes essential that PDF files are unencrypted and have accurate metadata. For example, PDF/A is regarded as the preferred format for digital text preservation in several recommendations including [1]. Encryption and poor metadata make the document less accessible and less searchable.

It has been pointed out that some PDF producing software create invalid files and research indicates that the quality of the PDF files is heterogeneous [2]. Termens et al. explored file types and PDF handling in two Spanish IRs and found many PDF files to be encrypted [3]. As the two repositories showed different tendencies for file handling, analyzing more repositories can potentially reveal a general approach for PDF handling.

The heterogeneity of PDF files is also an issue with PDF-embedded metadata. For example, reference manager services usually extract metadata embedded in PDF files. Serious problems can arise when the embedded metadata is different from the actual content. Sometimes PDF-embedded metadata can cause security issues. For example, Society for Industrial and Applied Mathematics releases guidelines for protecting referee personal information [4]. When PDF gets released on the Web, differences between published metadata and PDF-embedded metadata may create confusion.

This research aims to address the problem of the PDF's preservability by surveying PDF files in Japanese IRs. We investigate the following research questions:

RQ1 How many PDF files are prepared in preservable format?
RQ2 How many PDF files are encrypted in preservable manner?
RQ3 How many PDF files have good enough embedded metadata to go alone on the Web?

II. METHOD

We collected and analyzed the PDF files as follows:

1. **Collecting Metadata Records:** Metadata records were collected from IRs in Japan. In February 2019, we harvested metadata records from 582 IRs via ListMetadataRecords of Open Archives Initiative Protocol for Metadata Harvesting. All metadata was in the junii2 format that was prepared for JAIRO, which in turn is a federated search system

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

of Japanese IRs. This format contains “fulltextURL” element for URL of full text files. We collected 2,103,600 metadata records and 1,556,390 of these records had URLs. According to the statistics from JAIRO, this accounted for 74% of all metadata records in Japanese IRs in February 2019.

2. **Collecting PDF Files:** After collecting metadata records, PDF files were collected. We extracted 1,556,390 URLs from the fulltextURL element and tried to download all of them. Factors such as change of IRs’ URL and invalid URLs prevented downloading. A total of 1,509,767 files were downloaded. Most of the files were PDFs, but some were non-PDF file formats.

3. **Analyzing PDF Files:** Different types of information was extracted from PDF files using iText 7.1.0 library: PDF file information for RQ1, security information for RQ2, and metadata for RQ3. Some files failed to parse as PDF for various reasons including an invalid string in the PDF file. Finally, we analyzed 1,411,082 PDF files. TABLE I show the basic statistics of the files examined.

III. RESULTS

Only 0.9% of PDF files conformed to PDF/A (or PDF/B or PDF/UA), which is specialized for long-term preservation of electronic documents. 11.2% were tagged (structured) PDFs making them accessible to the visually impaired.

As shown in TABLE III, 30.5% of PDF files were encrypted. The "Printing not allowed" setting prevents open access. The "Screen Readers not allowed" setting prevents visually impaired people from extracting text from PDF file. Furthermore, encrypted PDF files cannot be converted to any other format in future.

To allow a PDF file to be distributed separately from the metadata of the IR, it is desirable that the PDF file have enough and accurate embedded metadata. Many PDF files did not have metadata embedded in “Document Information” part (48.9% had creator, 17.9% had title, and 1.5% had key words). In contrast, 35.7% of PDFs had creator metadata embedded in XMPMetadata part. Various types of PDF producing software were used as in a previous study [2]. Even if the metadata was embedded, the metadata was not necessarily accurate. Some PDF files still had their original file’s metadata and had not been updated to match the current file’s metadata. However, PDF files created by academic commercial publishers like Elsevier contained a lot of rich and accurate metadata.

IV. SUMMARY

The answers to our research questions are as follows:

RQ1 Few PDF files were made in a format that was appropriate for long-term preservation.

RQ2 30.5% of PDF files are encrypted and they might prevent future reuse.

RQ3 Many PDF files did not have sufficient metadata for distributing them separately from the metadata of the IRs.

Our results show that PDF files in Japanese IRs have several serious problems for their long-term preservation.

TABLE I
Basic Statistics

Metadata records in Japanese IRs	2,103,600
Full text URLs	1,556,390
Downloaded files	1,509,767
Downloaded PDF Files	1,509,470
Parsed PDF Files	1,411,082

TABLE II
PDF File Information for Long-Term Preservation

PDF/A	0.9%
Tagged PDF	11.2%

TABLE III
PDF Security Information

Encrypted	30.5%
Printing not allowed	0.6%
Screen Readers not Allowed	1.3%

TABLE IV
PDF Document Information and XMPMetadata

Creator in PDF Document Information	48.9%
Title in PDF Document Information	17.9%
Keywords in PDF Document Information	1.5%
Creator in PDF XMPMetadata	35.7%

REFERENCES

[1] Smithsonian Institution Archives. "Recommended Preservation Formats for Electronic Records." Internet: https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records [Mar. 16, 2019].

[2] H. Helin, K. Koivunen, J. Kylander, and J. Lehtonen, “402.2 PDF Mayhem: Is Broken Really Broken?,” 27-Dec-2018. [Online]. Available: osf.io/fzxc9.

[3] M. Termens, M. Ribera and A. Locher. "An analysis of file format control in institutional repositories", *Library Hi Tech*, Vol. 33, pp.162-174, 2015.

[4] Society for Industrial and Applied Mathematics. "Protecting Referee Personal Information." Internet: https://www.siam.org/Publications/Journals/Related/Journal-Policies/Detail/protecting-referee-personal-information, [Mar. 16, 2019].