CA2004

日本の機関リポジトリにおけるPDFファイルの長期保存とアクセシビリティ

安形　輝*
宮田洋輔†
池内　淳‡

1. はじめに

学術情報流通においてオープンアクセス環境が進展し、多くの研究成果がウェブ上においてPDF形式で公開されるようになっている。オープンアクセスを実現する手段にはオープンアクセスジャーナルなどさまざまなものがあるが、その一つの手段として機関リポジトリがある。機関リポジトリは、大学や研究機関において組織に属する研究者による研究成果を公開するためのサービスであり、日本において2020年9月時点で700件近く設置運用されている(1)。

機関リポジトリに登録される学術論文等の多くはPDF形式で作成されている。PDF（Portable Document Format）は、デバイスやOSに依存することなく表示・印刷可能なファイル形式であり、文書のレイアウトを保持できるため紙の文書との親和性が高いこと、テキスト情報だけではなく画像などのマルチメディアに対応していること、セキュリティレベルを自由に設定できることなどの特徴から、ネットワーク上の文書形式として広く利用されている。長期保存やアクセシビリティ等の観点からPDFの適正を図る上で重要であるのは次の4点である。

（1）PDF/Aであるか否か
（2）暗号化の有無
（3）ファイルの品質
（4）メタデータの品質

PDF形式の規格の一つであるPDF/Aは、ISO 19005(2)として国際規格となっており、たとえば、米・スミソニアン協会アーカイブの文書保存ガイドライン(3)をはじめとして、PDFファイルの長期保存に適した規格として推奨されている(4)。同規格では、要求要件や禁止要件などの仕様が厳密に定められており、メタデータやフォントを埋め込むことによって、将来的にもレイアウトなどが変わらず表示できるようになっている。また、PDF/UAはユニバーサルアクセシビリティ（Universal Accessibility）、つまり視覚障害者などであってもアクセスしやすいPDF形式の規格といえる。文書構造などをメタデータに埋め込むことが定められており、ISO 14289(5)として国際規格となっている。

PDFファイルを公開する際にあまり意識されていないことであるが、長期保存やアクセシビリティの視点から重要なのは、そのファイルを暗号化せず、正確なメタデータを持たせることである。暗号化と不適切なメタデータは、文書のアクセス性と検索性を低下させる可能性がある。例えば、不用意にセキュリティレベルを上げることで、文字の機械的な抽出が行えず、視覚障害者が文書を利用できないといった問題も起こり得る。

また、PDF作成ソフトの中にはPDFの仕様に合致しない不適切なファイルを作成するものがあることが指摘されており、PDFファイルの品質が均一ではないことが海外の調査で明らかになっている(6)。Termensらは、スペインの2つの機関リポジトリでファイルタイプとPDFのセキュリティを調査し、多くのPDFファイルが暗号化されていることを明らかにした(7)。この調査では調査対象が2つの機関リポジトリであるが、ファイルの取り扱いに異なる傾向が見られた。

PDFファイルの問題として、埋め込まれたメタデータが不適切である場合もある。例えば、PDFファイルを対象にした検索サービスではPDFファイルのメタデータをPDFファイルとは別に登録する場合もあるが、PDFファイルに埋め込まれたメタデータを抽出して、それらをインデックスに登録する場合もある。後者のようなサービスでは埋め込まれたメタデータが実際のコンテンツと異なる場合、検索ができなくなる問題が生じてしまう。また、PDFファイルに埋め込まれた作成者に関するメタデータが査読の匿名性を暴いてしまう問題を引き起こすことがある。そこで応用数理学会（Society for Industrial and Applied Mathematics：SIAM）は、査読における個人情報保護に関するガイドラインを発表しており、その中でPDFに含まれる個人情報を削除する手順を示している(8)。

このようなPDFファイルがウェブ上で公開された場合、公開されたメタデータとPDFに埋め込まれたメタデータの違いにより、混乱が生じる可能性がある。そこで、筆者らは日本の機関リポジトリにおけるPDFファイルの現状を明らかにし、PDFファイルの保存性の問題の解決策を検討することを目的とした調査を行った。

本稿ではその結果を紹介する。なお、本稿はiPRES2019での発表を再構成したものである(9)。

2. 調査手法

日本の機関リポジトリのPDFファイルの収集と分析

* 亜細亜大学国際関係学部
† 慶應義塾大学文学部
‡ 筑波大学図書館情報メディア系

は以下のように行った。

2.1 メタデータの収集

2019年2月に、OAI-PMHのListRecordsを介して、582の機関リポジトリからメタデータを収集した。すべてのメタデータは、日本の機関リポジトリの統合検索システムであるJAIRO（調査当時、現IRDB）のために用意されたjunii2形式である。この形式には、全文ファイルのURLを示す“fullTextURL”要素が含まれている。210万3,600件のメタデータを収集し、そのうち155万6,390件が全文ファイルのURLを持っていた。当時のJAIROに登録されていたメタデータのうち、74%となっている。

2.2 PDFファイルの収集

メタデータ・レコードを収集した後、PDFファイルを収集した。2.1でダウンロードしたメタデータから155万6,390件のURLを抽出し、その全てのダウンロードを試みた。ダウンロードできたファイルは150万9,767件で、ファイルヘッダの識別情報からはそのほとんどがPDFであることがわかったが、中にはPDF以外のファイル形式も含まれていた。

2.3 PDFファイルの分析

PDFファイルを操作できるJava版のiText 7.1.0ライブラリを用いて、ダウンロードできたPDFファイルから、セキュリティ情報などのメタデータを抽出した。なお、PDFファイルの中には、不正な文字列が含まれているなどの理由で、PDFとして解析できなかったファイルもあった。最終的に、141万1,082件のPDFファイルを解析した。表1に、収集したファイルの基本統計を示した。

表1　PDFファイルの基本統計

項　目	件　数
機関リポジトリから収集したメタデータ・レコード	210万3,600件
全文ファイルのURL	155万6,390件
ダウンロードしたファイル	150万9,767件
そのうち、PDFファイル	150万9,470件
そのうち、解析できたPDFファイル	141万1,082件

3. 調査結果

日本の機関リポジトリで公開されたPDFファイルのうち、電子文書の長期保存に特化したPDF/Aに準拠したものは、表2でまとめたように、0.9%であり、非常に少ないことがわかった。PDF/UAに準拠したものはさらに少なく9件（0.0006%）であり、ほぼこのフォーマットが採用されたファイルがないことがわかった。全体の11.2%はタグ付き（構造化）PDFファイルで、視覚障害者が読み上げソフトを使用する際に、利用しやすいものとなっていたが、こちらへの対応も全体から見ると少ないと指摘できる。

また、表3に示すように、30.5%のPDFファイルが暗号化されていた。暗号化設定のうち「印刷を許可しない」設定は、印刷して読みたい読者の読み方を制限する意味で完全なオープンアクセスといえるかは疑問が残る。また「スクリーンリーダーを許可しない」の設定は、視覚障害者がPDFファイルからテキストを抽出することを妨げるため、合理的配慮をしていないといえる。また、暗号化されたPDFファイルは、パスワードが明らかでない場合、将来的に新たなPDF形式や他のファイル形式に変換することができないという点で長期保存に向かないといえる。

PDFファイルはかならずしも機関リポジトリからダウンロードされる訳ではなく、メタデータと切り離して配布・流通することも想定される。その場合、PDFファイル自身に十分かつ正確なメタデータが埋め込まれていることが望ましい。しかし、表4で示したように、多くのPDFファイルは「文書情報」部分に必要なメタデータが埋め込まれていなかった（埋め込まれていたのは、作成者が48.9%、タイトルが17.9%、キーワードが1.5%）。拡張メタデータ領域であるXMPMetadata（Extensible Metadata Platform Metadata）に作成者のメタデータが埋め込まれているPDFは35.7%だった。メタデータに登録されていたPDF作成ソフト名に関しては、Helinらの先行研究[10]の結果と同様に、さまざまなものが使用されていることがわかった。メタデータが埋め込まれていたとしても、そのメタデータで作成者、タイトルおよび責任表示の誤っていたものも確認された。PDFファイルの中には、学会などが論文投稿用に配布した元のテンプレートファイルのメタデータが実際の作成者によって更新されず、そのまま残っているものもあった。一方で、管見の限りであるが、Elsevier社のような学術的な商業出版社が作成したPDFファイルには、メタデータの多くの項目に正確な情報が多い印象を受けるのとは対照的であった。

表2　PDFファイルの種類

PDFファイルの種類	割　合
PDF/A	0.9%
PDF/UA	0.0006%
タグ付き（構造化）PDF	11.2%

表３　PDF ファイルのセキュリティ設定

セキュリティ設定	割　合
何らかの暗号化がされていたもの	30.5%
印刷を許可しない	0.6%
スクリーンリーダーを許可しない	1.3%

表４　PDF ファイルに埋め込まれたメタデータ

メタデータ	割　合
PDF 文書情報に「作成者」あり	48.9%
PDF 文書情報に「タイトル」あり	17.9%
PDF 文書情報に「キーワード」あり	1.5%
PDF XMPMetadata に「作成者」あり	35.7%

4. まとめ

今回行った調査の結果から、日本の機関リポジトリにおいて公開された PDF ファイルについては、(1)PDF/A という長期保存に適した規格で作成された PDF ファイルはほとんどなかったこと、(2)30.5%の PDF ファイルが暗号化されており、アクセシビリティの視点から問題があり、将来的に他の形式への変換が阻害されていること、(3)多くの PDF ファイルは、機関リポジトリのメタデータから独立して流通する際の十分なメタデータが埋め込まれていなかったこと、の三点が明らかになった。

上記の結果から日本の機関リポジトリに含まれる PDF ファイルの多くは、長期保存やアクセシビリティの視点からはさまざまな問題を抱えているといえる。オープンアクセスを実現するために機関リポジトリで組織に所属する研究者の研究成果を PDF ファイルとして公開する場合に、単に受け取った PDF ファイルを置くだけでは十分とは言えない。長期保存やアクセシビリティを保障するためには、個々の PDF ファイルについて、PDF/A の規格に準拠したファイルとすること、不必要なセキュリティを設定しないこと、機関リポジトリにメタデータを登録するだけでなく正確で十分なメタデータを PDF ファイル自身に埋め込むことなどが重要だといえる。

学術文献のアクセシビリティについては「視覚障害者等の読書環境の整備の推進に関する法律」(CA1974 参照）を実現するための「視覚障害者等の読書環境の整備の推進に関する基本的な計画」(E2307 参照）[(11)]においていくつかの指摘がされている。まず、「〔全国の大学及び高等専門学校の附属図書館が保有するアクセシブルな書籍等の所在情報を共有するための〕リポジトリやデータベース等で公開される学術論文等について、視覚障害者等のアクセシビリティの向上に努める（括弧内筆者）」とある。また、「国立国会図書館において、学術文献の録音資料やテキストデータの製作を促進する」という記述に基づき、国立国会図書館が学術文献の視覚障害者等用テキストデータの図書館等からの製作依頼の受付を開始した[(12)]。また、点字図書館や一部の公共図書館でもテキストデータの製作を行っている。ただし、OCR（光学文字認識）での抽出には精度の問題があり、校正作業でのコストがかかることから OCR をせずとも PDF ファイルからテキストデータが直接的に抽出できる状態での公開が望ましい。視覚障害者等の読書環境の整備は国として推進している施策であり、今後ますます重要になってくると予想される。

(1) “機関リポジトリ一覧”. 学術機関リポジトリ構築連携支援事業. https://www.nii.ac.jp/irp/list/, (参照 2021-07-08).
(2) ISO 19005-1:2005. Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1). https://www.iso.org/standard/38920.html, (accessed 2021-07-08).
(3) “Recommended Preservation Formats for Electronic Records”. Smithsonian Institution Archives. https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records, (accessed 2021-05-08).
(4) Digital Preservation Team. PDF Format Preservation Assessment Part 2: PDF/A Profile. BL. 2019-06-30. https://wiki.dpconline.org/images/2/22/PDFA_Assessment_v1.0.pdf#page=8, (accessed 2021-07-08).
(5) ISO 14289-1:2014. Document management applications — Electronic document file format enhancement for accessibility — Part 1: Use of ISO 32000-1 (PDF/UA-1). https://www.iso.org/standard/64599.html, (accessed 2021-07-08).
(6) Helin, H.; Koivunen, K.; Kylander, J.; Lehtonen, J. “402.2 PDF Mayhem: Is Broken Really Broken?”. 15th International Conference on Digital Preservation iPRES 2018, Boston, USA, 2018-09-27, Center for Open Science. https://doi.org/10.17605/OSF.IO/FZXC9, (accessed 2021-05-08).
(7) Termens, M.; Ribera, M.; Locher, A. An analysis of file format control in institutional repositories. Library Hi Tech. 2015, vol. 33, no. 2, p. 162-174.
(8) “Protecting Referee Personal Information”. Society for Industrial and Applied Mathematics. https://www.siam.org/Publications/Journals/Related/Journal-Policies/Detail/protecting-referee-personal-information, (accessed 2021-05-08).
(9) Agata, T.; Miyata, Y.; Ikeuchi, A. “Long-term Preservation of Pdf Files in institutional Repositories in Japan”. 16th International Conference on Digital Preservation iPRES 2019, Amsterdam, The Netherlands. 2019 Center for Open Science. https://osf.io/xrwzq/, (accessed 2021-05-08).
(10) Helin. op. cit.
(11) “「視覚障害者等の読書環境の整備の推進に関する基本的な計画」の決定について”. 文部科学省. 2020-07-14. https://www.mext.go.jp/b_menu/houdou/mext_00265.html, (参照 2021-05-08).
(12) “学術文献の視覚障害者等用テキストデータの図書館等からの製作依頼を受け付けます”. 国立国会図書館. 2021-04-01. https://www.ndl.go.jp/jp/news/fy2021/210401_02.html, (参照 2021-05-08).

［受理：2021-08-10］

Agata Teru
Miyata Yosuke
Ikeuchi Atushi

Long-term Preservation of PDF Files in Institutional Repositories in Japan