タグ付きPDFとは何か?

sample

アンテナハウス株式会社

アンテナハウスCAS電子出版
CAS-UB
Cloud Authoring Service for Universal Book

はじめに

タグ付き PDF という言葉を見かけるようになりましたが、タグ付き PDF とは一体どういったものなのでしょうか。

本書ではタグ付き PDF とはなにか？　について簡単に説明します。

なお、本書は EPUB アクセシビリティの機能の 1 つである「ページ分割マーク」の CAS 記法マークアップと EPUB3 のサンプルとして用意されたものです。

目　　次

第1章 タグ付きPDFとはなにか

タグ付きPDFは、内部に文書構造を指定するタグを付与したPDFのことである。タグ付きPDFでは、テキストや画像などのコンテンツをマークで囲ってタグを付けるとともに、文書の階層や表などの構造を表すタグを追加する。そして、構造を表すタグとコンテンツを表すタグを使ってタグのツリー構造（タグツリー）を構築する。

文書の各部分に、部・章、見出し、引用、箇条書き、表などの構成上・意味上の役割を与えることを文書の構造化という。文書の構造化により、読み手が意味をより理解し易くなる。

タグ付きPDFは必須機能では無くオプション機能であり、現在、作成されているPDFの大部分はタグ付きPDFではない。これは次の理由による。PDFは、オフィスソフトなどのアプリケーションで文書を編集した結果を、プリンタで印刷する操作で作成するのが一般的である。印刷した文書は、主に、章・節などの区切り、版面内での配置、文字の大きさなどの視覚表現で構造を付けている。そこで、人間が文書を視覚的に読むだけならPDFの内部に文書構造を指定するタグを持たせる必要はないのである。

タグ付きPDFが有用になるのは、PDFを視覚的に読むときではなくPDF内部のデータを別の目的で使うときである。

第2章 タグ付きPDFでできること

タグ付きPDFの仕組みを利用してできることについて説明する。

2.1 PDFの内容の読み上げ

PDFをコンピュータで読み上げるときは内部に保存されているテキストを読む。しかし、PDFの内部に保存されているテキストの並び順は、文章の意味的な繋がりと一致しているとは限らない。タグ付きPDではタグツリーを辿るとコンテンツを読み上げる順序になる。

また、印刷では、柱やページ番号のようにナビゲーションのための情報や、本文領域の上や下の罫線、本文と脚注の間の罫線、テキストボックスの枠線や背景などの修飾的情報も多い。こうした修飾的情報は音声で読み上げるときは必要ない。タグ付きPDFは修飾的情報にはArtifactタグを付け、タグツリーには登録しない。

2.2 PDFを変換して再利用する

タグ付きPDFではセクション・見出し・表・段落などの論理的な構造をPDFに追加できる。この構造を利用すればPDFからオフィスなどの編集用ファイルに戻すときに、より適切な変換ができる。また、PDFからHTML（Webページ）に変換するときも、見出し・表・箇条書きのような構造を使って、より良い結果を得ることができる。

2.3 PDF のリフロー表示

タグ付き PDF で決めている標準タグは、Web ページを記述するための HTML に類似している。タグ付き PDF に準拠するリーダーは、印刷のためのレイアウトで表示するだけでなく、PDF 内部に設定されているタグを使って表示もできる。これにより、あたかも HTML をスマホの画面上にブラウザで表示するように、PDF 表示でも画面の端でテキストを折り返して（リフロー）表示ができる（CAS-UB ブログ「PDF のリフロー表示。タグ付き PDF とタグの付いていない PDF の比較。」（p. 13））。

2.4 アクセシビリティサポート

PDF のアクセリビリティサポートの中核は既に説明した読み上げ順序や論理構造である。その他にタグ付き PDF の仕組により、①文書の言語指定、②イメージ・数式などテキストに翻訳できない項目の代わりに読み上げるための代替テキストの設定、③略語・頭字語などに対する展開語（例えば ISO という頭字語に International Organization for Standardization）の指定ができる。

第3章 タグ付きPDFの標準タグと属性

タグ付きPDFでは標準タグの種類を決めている。標準以外のタグを使うこともできるが、その場合は標準的なタグとの対応関係が分かるようにしなければならない。また、属性の標準も決めている。属性の多くはリフロー表示のためのレイアウト属性とPDFの内容を他の形式に変換する時に参照されるものである。箇条書きのラベルと表のセルに関する属性もある。次にどのような標準タグがあるかを示す。

3.1 グループ化のためのタグ

他のタグをグループ化するために使うタグである。タグツリーはトップレベルのタグを一つだけもつ。完全なドキュメントのときトップレベルのタグは、Documentとする。ドキュメントの断片のときPart、Art、Sect、Divのどれか一つをトップレベルのタグとするのが良い。

他のグループ化のタグには、ブロック引用（BlockQuote）、キャプション（Caption）、目次（TOC）、目次項目（TOCI）、インデックス（Index）がある。

3.2 ブロックレベルのタグ

段落（P）、見出し（H、H1～H6）、箇条書き（L、LI、Lbl、LBody）のようにドキュメントの行を積み重ねていく方向に配置するテキストやその他の内容領域を示す。

3.3 テーブルのためのタグ

テーブル（Table）タグは、ブロックレベルのタグである。下位のタグとしては、テーブル行（TR）、テーブルヘッダーセル（TH）、テーブルデーターセル（TD）、テーブルヘッダー（THead）、テーブルボディ（TBody）、テーブルフッター（TFoot）がある。これらはテーブルの内部を構造化するタグである。

3.4 テーブルの標準属性

テーブルは行と列から構成されるが、幅広い表の構造を表現するにはセル結合などのための機能も必要である。このためにRowSPan（自然数）、ColSpan（自然数）、Headers（配列）、Scope（名前）、Summary（文字列）といった属性が使える。

3.5 行内のためのタグ

文書の中のテキストの一部を表すタグである。行内で文字の進行方向に積み重ねる。スパン（Span）、引用（Quote）、ノート（Note）、参照（Reference）、目録エントリー（BibEntry）、コード（Code）、リンク（Link）、注釈（Annot）がある。

3.6 イラストのためのタグ

イラストタグは図（Figure）、数式（Formula）、フォーム（Form）のどれかである。イラストが文書内の段落の一部になっていることがある。このような場合はFigureタグを使って表現する。

第4章 タグ付きPDFの例

次のような一ページの簡単なPDFをタグ付きPDFにする例を示す。

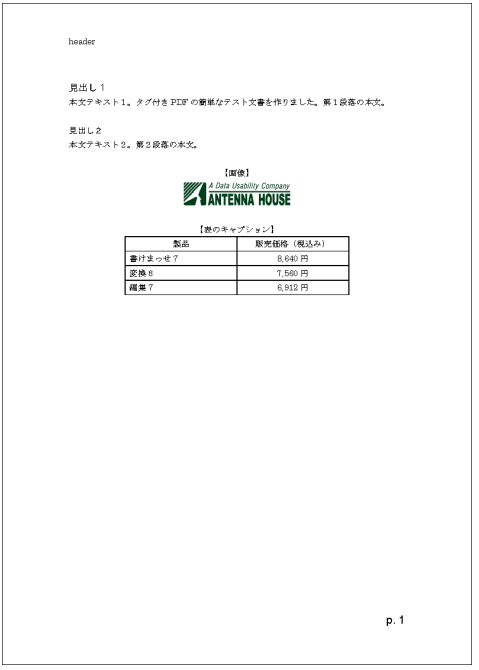

図4.1 PDFの例

この文書は、見出し1とその本文、見出し2とその本文、画像のキャプションと画像、表のキャプションと表、という順序になって

いる。本文領域の上にヘッダー、下にページ番号が印刷されている。

PDF のコンテントには次のようにマークアップする。

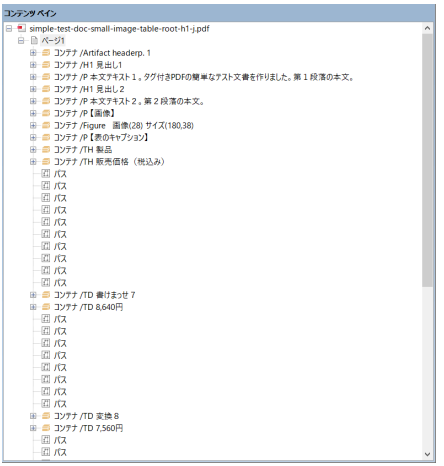

図 4.2 PDF コンテントへのマークアップの例

タグ付き PDF でタグツリーをどのように作るかは一意では無く自由度が高い。ここではルートタグは Sect とし、見出し 1、その本文、見出し 2、その本文、画像のキャプション、画像（イメージ）、表とキャプション、がすべて兄弟関係とする。つまり、第一階層がすべてルートタグ（Sect）の子供であると考えてタグツリーを作ると次のようになる。

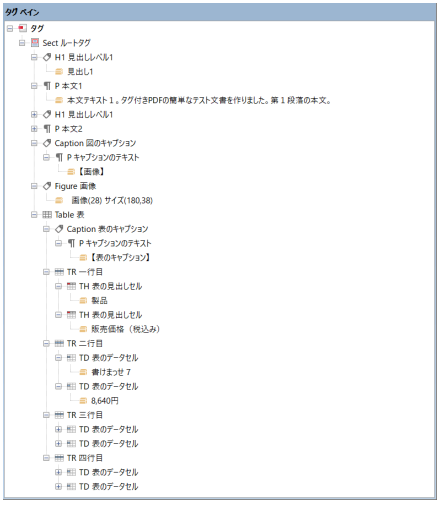
タグ ペイン
タグ
Sect ルートタグ
H1 見出しレベル1
見出し1
P 本文1
本文テキスト１。タグ付きPDFの簡単なテスト文書を作りました。第１段落の本文。
H1 見出しレベル1
P 本文2
Caption 図のキャプション
P キャプションのテキスト
【画像】
Figure 画像
画像(28) サイズ(180,38)
Table 表
Caption 表のキャプション
P キャプションのテキスト
【表のキャプション】
TR 一行目
TH 表の見出しセル
製品
TH 表の見出しセル
販売価格（税込み）
TR 二行目
TD 表のデータセル
書けまっせ 7
TD 表のデータセル
8,640円
TR 三行目
TD 表のデータセル
TD 表のデータセル
TR 四行目
TD 表のデータセル
TD 表のデータセル

図 4.3 タグツリーの例

第5章 タグ付きPDFの採用は進む

5.1 官公庁・行政での採用

欧米の政府関係機関においては、タグ付きPDFはPDFアクセシビリティの重要な要素として普及している。それに対して、日本ではタグ付きPDFについてはあまり注目されてこなかった。しかし、2016年4月より施行された障害者差別解消法では、官公庁・行政機関は、障害者より要求があったときは実施に伴う負担が過重でない範囲で情報をアクセシブルにすることが義務付けられている。こうしたことで日本でも官公庁や行政はタグ付きPDFの採用が始まっている。

5.2 プロファイル仕様への採用

PDFの全機能はあまりにも多い。そこで、利用者の立場から用途を絞った仕様が提案されている。こうした機能の使い方を定める仕様をプロファイル仕様という。タグ付きPDFは、PDFのプロファイル仕様である長期保存（PDF/Aファミリー）やPDFのアクセシビリティー（PDF/UA）の一部として採用されている。

参考資料

「PDF のリフロー表示。タグ付き PDF とタグの付いていない PDF の比較。」 <Web http://blog.cas-ub.com/?p=6581>

CAS-SUPPORT

アンテナハウスの電子書籍制作サービス CAS-UB のサポートチーム

タグ付き PDF とは何か？

2017 年 9 月 29 日　初版

著　　者　CAS-SUPPORT
発 行 者　CAS 電子出版
発 行 所　アンテナハウス株式会社
住　　所　東京都中央区東日本橋 2 丁目 1 番 6 号
電話番号　03-5829-9021
Ｗ Ｅ Ｂ　http://www.cas-ub.com/support/