PDFテキスト抽出の基本、問題点、実践的解決策

2020年8月25日　16:00～17:20

アジェンダ

テーマ：　PDFテキスト抽出の基本、問題点、実践的解決策

プレゼンター：小林　徳滋　（koba@antenna.co.jp）

時間	見出し	内容
16:00 ～ 16:20	PDFのテキスト抽出の基本	PDFで文字を表示する仕組み、PDF内のテキストの扱いについて簡単に解説します。
16:20 ～ 16:40	PDFからテキストをコピー&ペーストするときの問題	PDFからテキストをコピー&ペーストする際に経験することがある、様々なトラブルの例を紹介し、トラブルが発生する原因について説明します。
16:40 ～ 17:00	PDFのテキスト抽出のソリューション	現在、アンテナハウスで開発中のPDF高度テキスト抽出アプリケーションを紹介します。簡単なデモを予定しています。
17:00 ～	質疑	チャットでの質疑応答と必要に応じて補足説明します。

※時間は目安です。進行の具合によって多少前後しますのであらかじめご了承ください。

はじめに：作り方によるPDFの大分類

ボーンデジタルPDFと電子化PDF

- ボーンデジタルPDF
  - オフィスソフト・DTPソフトなどのアプリで出力したPDF、帳票ソフトやレポート作成ソフトから出力したPDF、など
  - ボーンデジタルPDFは**フォントを使って文字が表示**される
- 電子化PDF
  - 紙をスキャンした画像から作るPDF、デジタル写真からPDF変換で作成するPDF
  - 電子化PDFでは**文字はビットマップ画像**で表される。文字の表示にフォントを使わない
  - 電子化PDFのテキスト抽出は、PDFを画像に戻してOCR処理で文字を認識する。OCR文字認識の精度は100%ではないため、誤った文字になる可能性がある

※今日の話題は、主にボーンデジタルPDFからのテキスト抽出に関する内容です。

ボーンデジタルPDFからのテキスト抽出

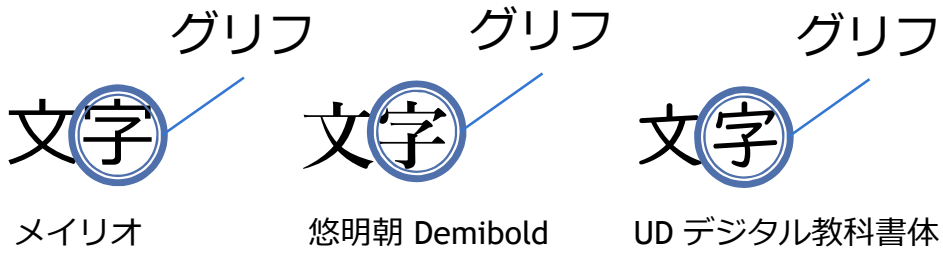
- たいていのボーンデジタルPDFからは、 PDFファイル内にある、文字を表示するためのデータからテキスト抽出できる

  こちらが今日のトピックです。

- そのメリット
  - OCRと違って文字コードの誤認識がない
  - 文字の位置を精密に取得できるので位置情報を使える
- そのデメリット
  - PDFのセキュリティで、テキストのコピーが許可されている必要がある
  - 文字コードを取り出せないケースがある
  - PDFの作り手によって内部が千差万別なため、いろいろなパターンのPDFに対応するための開発工数が大きい

フォント、文字、グリフ
フォントで文字を表示するとは

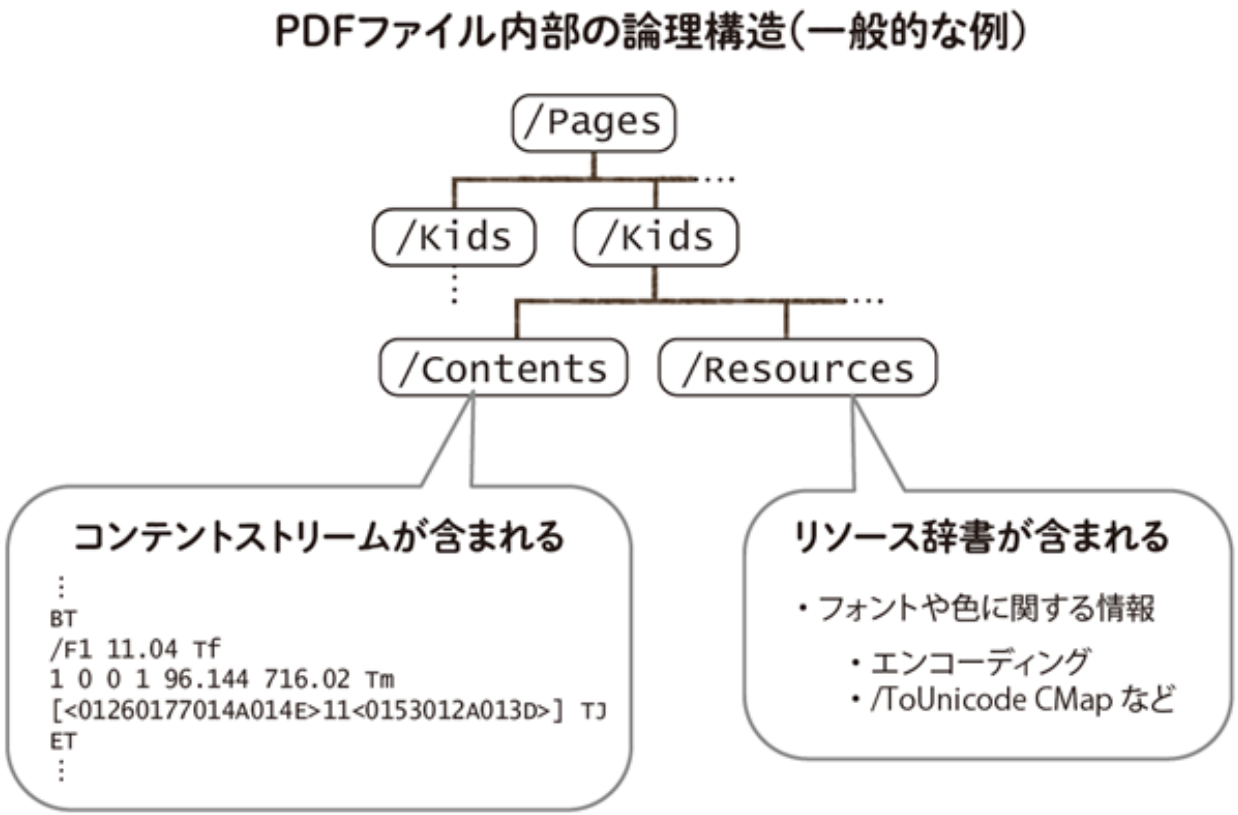
- 文字の形（字形）を表す絵のことを**グリフ**という（用語）
- フォントの中にはグリフを描画するデータがある
- フォントに文字の識別子を入力して、グリフを探す
- フォントのエンコーディングは、文字の識別子からフォント内のグリフへの対応関係を示す



※同じ文字でも、フォントによってグリフが異なる

PDFのテキスト表示の仕組み

- PDFのテキストとはページの特定位置に描かれる文字のこと
- 表示される文字の識別子と表示位置などのデータは、PDFのページ内のコンテントストリームに置かれている
- フォントの情報はフォントリソースに置かれている
- PDFリーダーは、コンテントストリーム内の文字の識別子とフォントのエンコーディングを元に、フォントの中のグリフを特定し、指定された位置に文字を印刷・表示する
- コンテントストリームはページの中で細切れになっている
- PDF仕様（ISO 32000-1）は文字の識別子からグリフを探して表示する仕組みについて決めている
  - ISO 32000-1:2008　「第9章　テキスト」を参照



プログラマーから見たPDFファイル
https://www.antenna.co.jp/pdf/reference/pdftext.html

PDFの表示

PDF Text Extractor
PDFテキスト抽出

← Times New Roman

← 悠明朝

フォント埋め込みなしで作成したPDFの内部

〇コンテントストリーム

コンテントストリームに改行コードはない

```
BT
/F0 10.56 Tf 85.08192 729.83997 Td [(P) -1(D) -5(F) -1( ) 34(T) 66(e) 1(xt) 5( )
11(E) -2(xt) 5(r) 3(a) 1(c) 1(t) 5(or) ] TJ
/F1 10.56 Tf 0.00004 -17.88 Td [<0050>
4<0044> 4<0046> -250<30C630AD> 12<30B930C8> 11<62BD51FA> ] TJ
ET
```

使用フォント、フォントサイズ

行頭の位置

表示する文字の識別子の並び

〇フォントリソース

```
<</Font <</F0 9 0 R/F1 10 0 R>>/

9 0 obj
<</Type /Font/Subtype /TrueType/Name /F0/BaseFont
/TimesNewRomanPSMT/FirstChar 32/LastChar 120/Widths [250 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 722 611 556 0 0 0 0 0 0 0 0 556 0 0 0
611 0 0 0 0 0 0 0 0 0 0 0 0 444 0 444 0 444 0 0 0 0 0 0 0 0 0 500 0 0 333 0 278 0 0 0
500]/Encoding /WinAnsiEncoding/FontDescriptor 11 0 R>>
endobj

10 0 obj
<</Type /Font/Subtype /Type0/Name /F1/BaseFont /YuMincho-Regular/Encoding
/UniJIS-UCS2-H/DescendantFonts [12 0 R]>>
endobj
```

PDFの表示

PDF Text Extractor
PDFテキスト抽出

← Times New Roman

← 悠明朝

フォント埋め込みで作成したPDFの内部

コンテントストリームに改行コードはない

○コンテントストリーム

```
BT /F0 10.56 Tf 85.08192 729.83997 Td [<0033> -1<0027> -5<0029> -1<0003> 34<0037>
66<0048> 1<005B0057> 5<0003> 11<0028> -2<005B0057> 5<0055> 3<0044> 1<0046>
1<0057> 5<00520055> ] TJ
/F1 10.56 Tf 0.00004 -17.88 Td [<0033> 4<0027> 4<0029> -250<03C403AB>
12<03B703C6> 11<0BAA095C> ] TJ
ET
```

使用フォント、フォントサイズ

行頭の位置

表示する文字の識別子の並び

○フォントリソース

```
<</F0 9 0 R/F1 10 0 R>>

9 0 obj
<</Type /Font/Subtype /Type0/Name /F0/BaseFont
/JDTVFA+TimesNewRomanPSMT/Encoding /Identity-H/ToUnicode 12 0 R/DescendantFonts
[11 0 R]>>
endobj

10 0 obj
<</Type /Font/Subtype /Type0/Name /F1/BaseFont /JDTVFA+YuMincho-Regular/Encoding
/Identity-H/ToUnicode 16 0 R/DescendantFonts [15 0 R]>>
endobj
```

情報処理における文字とテキストデータとは

- コンピュータのプログラムで文字を取り扱うためには標準化された文字コードを使う。標準文字コードは大雑把には次のようなもの
  1. 印刷物などで実際に使われている文字を収集する。
  2. 字義が同じで字体が同一視できる文字を一つにまとめる。これを包摂する（Unification）と言う。文字はこのように抽象化される
  3. こうして作成した文字集合の構成要素に一定の番号（コードポイント）を与える。できたものを符号化文字集合と言う
  4. コンピュータで情報処理するときはコードポイントを一定の計算式でバイトのならびに変換して取り扱う。これが、情報処理におけるテキストデータであり、この変換方式を符号化方式と言う
- 符号化文字集合ではUnicodeが一般的。Unicodeをコンピュータで使うときの符号化方式はUTF-8、UTF-16などが使われる
- 日本の符号化文字集合としてJIS文字規格は幾つかある。JIS X0213:2012が最新のJIS文字規格

PDFテキスト抽出の内部処理

- PDFでのテキストはコンテントストリーム内にある文字識別子の並び（表示する文字）のことである
- 情報処理のテキストは、Unicodeなどで表されたデータ
- テキスト抽出では、次の二つの処理が必要
  1. PDF内の文字識別子の並びをUnicodeのテキストデータに変換する（文字コード変換）
  2. ページ上で文字が表示される位置を考慮しながら、テキストの流れを作る（文の流れの作成）

PDFからテキストを取り出すには

- ユーザーがPDFファイルからテキストを取り出すとき：
  1. PDFをリーダーで画面に表示する
  2. 画面上で、テキストを選択し、コピーする
  3. メモ帳などにペーストする
- ペーストされた結果を保存すると、形式的にはテキストデータになる

PDFテキストの コピー＆ペーストの問題例

- ペーストしたテキストが期待通りになっていないことが多い
- PDFからテキストをコピー＆ペーストしたとき起きる問題の実例のいくつかを紹介

１．文の順序が表示順と違う

- 画面に表示している順序とコピー＆ペーストしたテキストで段落の順序が異なる
- Officeソフトから作成したシンプルなレイアウトのPDFでも起きることがある（例１）
- DTPソフトで作成した複雑なレイアウトのPDFではかなり頻繁にみられる（例２）

例１

休業実績一覧表（様式特小第２号（小規模事業主用様式））

【記入要領】

１　「支給申請する１か月間（判定基礎期間）」

雇用調整助成金では、原則として１か月単位で休業の実績について確認し、それに基づいて支給がなされます。この休業の実績を判定する１か月単位の期間を「判定基礎期間」といいます。この単位になる１か月は、休業した事業所の毎月の賃金の締め切り日の翌日から、その次の締め切り日までの期間です。毎月の賃金の締め切り日がない場合などは、暦月（カレンダーの１日～月末日）となります。

支給申請のときは、この判定基礎期間を単位として、令和２年1月24日以降の期間であれば複数月分をまとめて申請することができます。その場合、この「休業実績一覧表」は、判定基礎期間ごとに作成する必要があります。

２　「従業員の数」

２か月を超えて使用される者（実態として２か月を超えて使用されている者のほか、それ以外の者であっても雇用期間の定めのない者及び２か月を超える雇用期間の定めのある者を含む。）であり、かつ、週当たりの所定労働時間が、当該事業主に雇用される通常の労働者と概ね同等（現に当該事業主に雇用される通常の労働者の週当たりの所定労働時間が４０時間である場合は、概ね４０時間である者をいう。ただし、労働基準法（昭和２２年法律第４９号）の特例として、所定労働時間がいまだ４０時間を上回っている場合は、「概ね同等」とは、概ね当該所定労働時間を指す。）である者をいいます。

３　「休業手当支払い率」

労働者の代表と休業の方法について約束するときに決めた休業手当の支払い率を、記入してください。

なお、休業するときに労働者に支払う休業手当の額は、通常支払っている賃金の６０％以上である必要があります。通常支払っている賃金と同じ額を支払っているときは、１００％と記入してください。

対象労働者ごとにちがう複数の支払い率がある場合は、最も多い労働者に適用している支払い率としてください。または、すべての支払い率の単純平均か、加重平均で計算した支払い率でもかまいません。

（例）支払い率が６０％の従業員５人、８０％の従業員２人、１００％の従業員３人の場合

最も多い従業員に適用している支払い率：６０％
単純平均：（６０＋８０＋１００）÷３種類＝８０（％）
加重平均：（６０×５ ＋ ８０×２ ＋ １００×３）÷１０人＝７６（％）　のうちいずれかを選択

丸一日休業した場合と、１日のうちの一部休業した場合で、ちがう支払い率としている場合は、加重平均で計算した支払い率としてください。

（例）丸一日休業したときの支払い率９０％で１０日休業し、
１日のうち一部休業したときの支払い率８０％で３日分（※）休業した場合　　※４で一日分換算した⑥の日数

加重平均：（９０×１０ ＋ ８０×３）÷１３日＝８８（％）

４　休業した事業所で、労働者が通常１日に働く労働時間数を記入してください。（就業規則や雇用契約書、労働条件通知書などに記載している労働時間数です。）

労働者ごとに所定労働時間がちがう場合は、最も多い労働者に適用している所定労働時間数としてください。

５　⑥欄は「④一部の時間帯休業した時間数」の合計を、４の時間数で割ることにより、何日分休業したことになるかを計算して、記入してください。小数点以下は切り上げて、整数で記入してください。

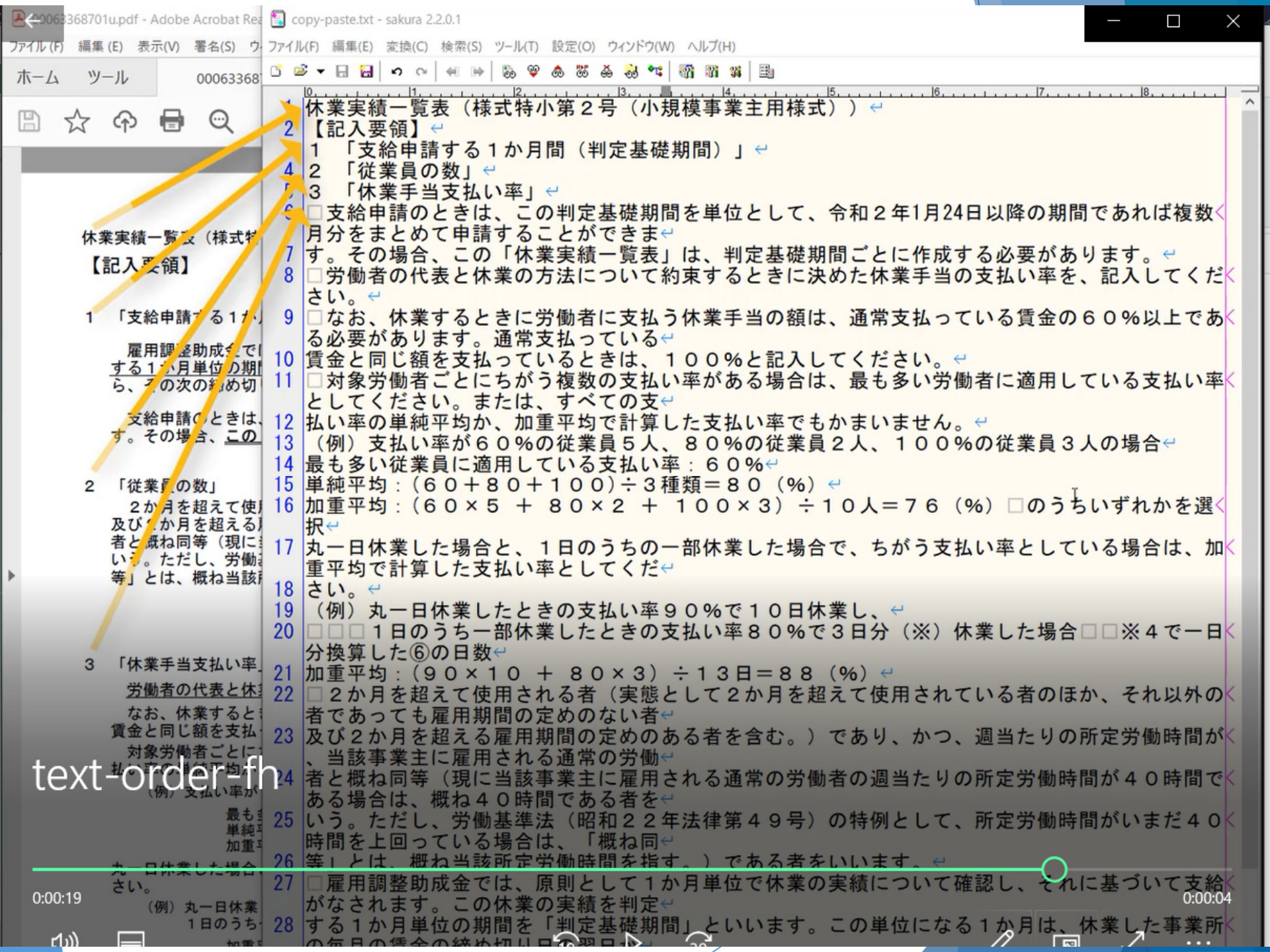
６　休業対象労働者ごとの休業実績一覧

解雇予告をされた者、退職願を提出した者、退職勧奨に応じた者は含めることができません（それぞれの時点より前までの休業についてであれば含めても可）。また、併給ができない他の助成金の対象労働者は含めることができません。

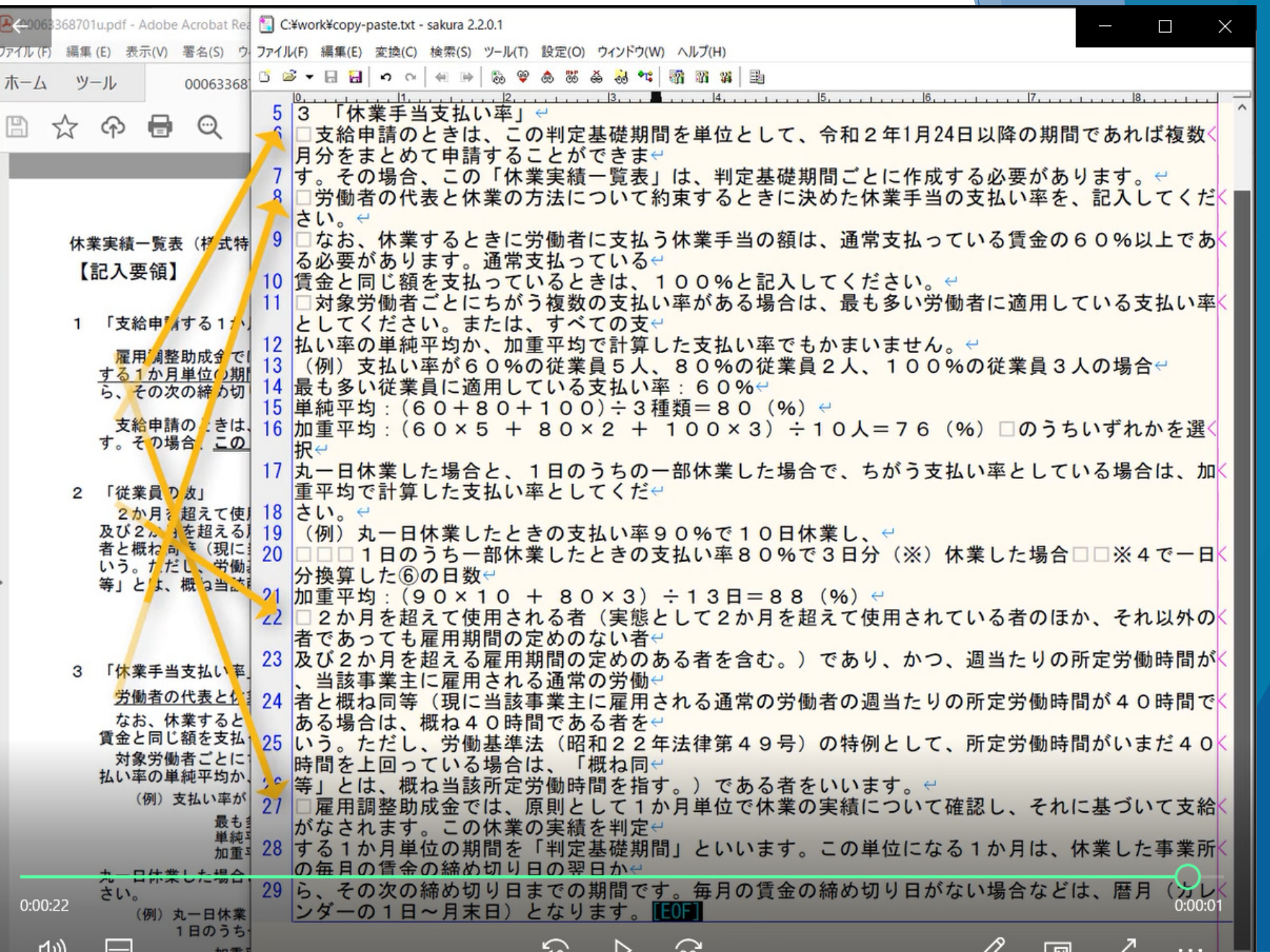
①	氏名	できれば、添付して提出する「休業手当の額がわかる書類（賃金台帳や給与明細など）」か「休業させた日や時間がわかる書類（出勤簿やタイムカードなど）」と同じ順番になるように記入してください。
②	雇用保険被保険者番号	対象労働者を雇用保険に加入させたときに発行される「雇用保険被保険者証」などに記載されている被保険者番号を記入してください。（※雇用保険に加入させていない従業員については、「緊急雇用安定助成金」というもうひとつの助成金の申請をしてください。）
③	１日休業した日数	丸一日休業した日数を記入してください。
④	１日のうち一部休業した時間数	通常の一日の営業時間のうち、一部休業した場合の時間数を記入してください。具体的には、短時間休業の時間（30分未満は切り捨て。例：１時間40分→1.5）数の合計を記入してください（合計欄は小数点以下切り上げ。）。
⑤	判定基礎期間の休業手当の額	「判定基礎期間」中の休業について、「休業手当支払い率」にもとづき、対象労働者ごとに支払う休業手当の額を記入してください。１日休業した場合（③）と１日のうち一部休業した時間数（④）の額をわける必要はありません。
	※③～⑤欄は、一覧に記載した対象労働者すべての数を合計して、合計欄に記入してください。	
⑦	休業延べ日数	③の合計（一覧に記載した対象労働者ごとの丸一日休業した日数の合計）と、⑥の日数の、合計日数を記入してください。

７　一覧表の下にある文章を確認し、記名押印または署名してください。また、休業に関する内容が事前に確約した内容であることについて労働者代表の方に確認してもらい、記名押印または署名をもらってください。

様式特小第２号
（新型コロナウイルス感染症関係）
（小規模事業主用様式）
（厚生労働省）



text-order-fh


C:¥work¥copy-paste.txt - sakura 2.2.0.1
ファイル(F) 編集(E) 変換(C) 検索(S) ツール(T) 設定(O) ウィンドウ(W) ヘルプ(H)
3 「休業手当支払い率」
□支給申請のときは、この判定基礎期間を単位として、令和２年1月24日以降の期間であれば複数月分をまとめて申請することができます。その場合、この「休業実績一覧表」は、判定基礎期間ごとに作成する必要があります。
□労働者の代表と休業の方法について約束するときに決めた休業手当の支払い率を、記入してください。
□なお、休業するときに労働者に支払う休業手当の額は、通常支払っている賃金の６０％以上である必要があります。通常支払っている賃金と同じ額を支払っているときは、１００％と記入してください。
□対象労働者ごとにちがう複数の支払い率がある場合は、最も多い労働者に適用している支払い率としてください。または、すべての支払い率の単純平均か、加重平均で計算した支払い率でもかまいません。
（例）支払い率が６０％の従業員５人、８０％の従業員２人、１００％の従業員３人の場合
最も多い従業員に適用している支払い率：６０％
単純平均：（６０＋８０＋１００）÷３種類＝８０（％）
加重平均：（６０×５ ＋ ８０×２ ＋ １００×３）÷１０人＝７６（％）□のうちいずれかを選択
丸一日休業した場合と、１日のうちの一部休業した場合で、ちがう支払い率としている場合は、加重平均で計算した支払い率としてください。
（例）丸一日休業したときの支払い率９０％で１０日休業し、
□□□１日のうち一部休業したときの支払い率８０％で３日分（※）休業した場合□□※４で一日分換算した⑥の日数
加重平均：（９０×１０ ＋ ８０×３）÷１３日＝８８（％）
□２か月を超えて使用される者（実態として２か月を超えて使用されている者のほか、それ以外の者であっても雇用期間の定めのない者及び２か月を超える雇用期間の定めのある者を含む。）であり、かつ、週当たりの所定労働時間が、当該事業主に雇用される通常の労働者と概ね同等（現に当該事業主に雇用される通常の労働者の週当たりの所定労働時間が４０時間である場合は、概ね４０時間である者をいう。ただし、労働基準法（昭和２２年法律第４９号）の特例として、所定労働時間がいまだ４０時間を上回っている場合は、「概ね同等」とは、概ね当該所定労働時間を指す。）である者をいいます。
□雇用調整助成金では、原則として１か月単位で休業の実績について確認し、それに基づいて支給がなされます。この休業の実績を判定する１か月単位の期間を「判定基礎期間」といいます。この単位になる１か月は、休業した事業所の毎月の賃金の締め切り日の翌日から、その次の締め切り日までの期間です。毎月の賃金の締め切り日がない場合などは、暦月（カレンダーの１日～月末日）となります。[EOF]
0:00:22
0:00:01

例２

特集 弾ける美味さ！ 信州中野のおいしいぶどう

日本一の産地を目指して―。味にこだわる信州中野のぶどう

武田政志（たけだまさし）さん（竹原）

わが家のぶどう栽培

両親と妻と私の4人で「巨峰」、「ナガノパープル」、「ピオーネ」、「シャインマスカット」、「クイーンニーナ」など8種類のぶどうを栽培しています。わが家は元々リンゴなどを栽培する農家でしたが、祖父の代からぶどうの栽培を始め、私はぶどう農家としては3代目になります。

現在は「JA中野市ぶどう部会」の部会長を務めており、県外へぶどうのセールスに行く機会も多くあります。お客さんの「おいしい」という笑顔を見ると嬉しいもので、来年も良いものを作らなければという意欲が湧いてきます。

中野市のぶどうの歴史

本市で本格的に巨峰の栽培が始まったのは昭和28年ごろだといわれています。ぶどうといえば「デラウェア」や「ナイアガラ」が主流だった当時、大粒で味が抜群に良い「巨峰」が登場しました。当初は栽培が難しく安定した生産ができなかったようですが、次第に栽培技術が発達し、安定生産が可能になったことで栽培農家が増え、中野市は日本有数の巨峰の産地となりました。

現在では、5月に出荷する超加温栽培、6～8月に出荷する加温・無加温栽培、9～11月に出荷する露地栽培、11月に出荷する抑制栽培、12～1月に出荷する冷蔵出荷と、5月から翌年1月までの長期間にわたり、味にこだわったぶどうを出荷しています。

仲間との支え合い

地域のぶどう農家には、産地の歴史を築いてきた先輩方だけでなく、若い後継者が入ってきて活気があり、とても心強く感じています。

畑で栽培技術の話をしたり、お互いのハウスのビニールを張る作業をしたりして、仲間同士支え合ってぶどう栽培をしています。昨年のことですが、突風でハウスのビニールが捲れ上がってしまったときに、園主が不在にもかかわらず近所の仲間たちが集まって直してくれたという出来事もありました。

中野のぶどうを全国へ

地元で採れた新鮮でおいしいぶどうを「おつかいもの」などに使っていただき、「信州中野のぶどう」を自慢の逸品として皆さんにも宣伝していただければと思います。

信州中野発！

おいしいぶどうができるまで

甘くて果汁がたっぷりのおいしいぶどうはどのように栽培されているのでしょうか？

巨峰の露地栽培の様子を、JA中野市の営農指導員の方に聞きました。

②発芽（4月）：春を迎え、ぶどうの木から若い新芽が顔を出します

①剪定（12月～2月上旬）：枝を切ることにより降雪によるぶどう棚の倒壊を防ぎます

⑨着色（8月）：巨峰は1粒ずつ着色が進んでいきます

⑩収穫（9月）：手間暇かけて育てたぶどうがついに収穫を迎えます

おいしいぶどうは、農家の皆さんの努力と愛情の結晶です！

▲JA中野市でぶどうの営農指導員を務める清水達哉さん

中野市
別冊 広報なかの vol.1
P.10
https://www.city.nakano.nagano.jp/docs/2015121800024/file_contents/1011.pdf


特集 弾ける美味さ！ 信州中野のおいしいぶどう

日本一の産地を目指して―。味にこだわる信州中野のぶどう

武田政志さん（竹原）

わが家のぶどう栽培

両親と妻と私の４人で「巨峰」、「ナガノパープル」、「ピオーネ」、「シャインマスカット」、「クイーンニーナ」など８種類のぶどうを栽培しています。わが家は元々リンゴなどを栽培する農家でしたが、祖父の代からぶどうの栽培を始め、私はぶどう農家としては３代目になります。

現在は「ＪＡ中野市ぶどう部会」の部会長を務めており、県外へぶどうのセールスに行く機会も多くあります。お客さんの「おいしい」という笑顔を見ると嬉しいもので、来年も良いものを作らなければという意欲が湧いてきます。

中野市のぶどうの歴史

本市で本格的に巨峰の栽培が始まったのは昭和28 年ごろだといわれています。ぶどうといえば「デラウェア」や「ナイアガラ」が主流だった当時、大粒で味が抜群に良い「巨峰」が登場しました。当初は栽培が難しく安定した生産ができなかったようですが、次第に栽培技術が発達し、安定生産が可能になったことで栽培農家が増え、中野市は日本有数の巨峰の産地となりました。

現在では、5月に出荷する超加温栽培、６～８月に出荷する加温・無加温栽培、９～ 11 月に出荷する露地栽培、11月に出荷する抑制栽培、12 ～ １月に出荷する冷蔵出荷と、5月から翌年１月までの長期間にわたり、味にこだわったぶどうを出荷しています。

仲間との支え合い

地域のぶどう農家には、産地の歴史を築いてきた先輩方だけでなく、若い後継者が入ってきて活気があり、とても心強く感じています。

畑で栽培技術の話をしたり、お互いのハウスのビニールを張る作業をしたりして、仲間同士支え合ってぶどう栽培をしています。昨年のことですが、突風でハウスのビニールが捲れ上がってしまったときに、園主が不在にもかかわらず近所の仲間たちが集まって直してくれたという出来事もありました。

中野のぶどうを全国へ

地元で採れた新鮮でおいしいぶどうを「おつかいもの」などに使っていただき、「信州中野のぶどう」を自慢の逸品として皆さんにも宣伝していただければと思います。

おいしいぶどうができるまで

②発芽（４月）：春を迎え、ぶどうの木から若い新芽が顔を出します

①剪定（12月～２月上旬）：枝を切ることにより降雪によるぶどう棚の倒壊を防ぎます

⑨着色（８月）：巨峰は１粒ずつ着色が進んでいきます

⑩収穫（９月）：手間暇かけて育てたぶどうがついに収穫を迎えます

信州中野発！

甘くて果汁がたっぷりのおいしいぶどうはどのように栽培されているのでしょうか？巨峰の露地栽培の様子を、ＪＡ中野市の営農指導員の方に聞きました。

おいしいぶどうは、農家の皆さんの努力と愛情の結晶です！

▲ＪＡ中野市でぶどうの営農指導員を務める清水達哉さん

別冊 広報なかの vol.1 10

特集
弾ける美味さ！
信州中野のおいしいぶどう

わが家のぶどう栽培

両親と妻と私の４人で「巨峰」、「ナガノパープル」、「ピオーネ」、「シャインマスカット」、「クイーンニーナ」など８種類のぶどうを栽培しています。わが家は元々リンゴなどを栽培する農家でしたが、祖父の代からぶどうの栽培を始め、私はぶどう農家としては３代目になります。

現在は「ＪＡ中野市ぶどう部会」の部会長を務めており、県外へぶどうのセールスに行く機会も多くあります。お客さんの「おいしい」という笑顔を見ると嬉しいもので、来年も良いものを作らなければという意欲が湧いてきます。

中野市のぶどうの歴史

本市で本格的に巨峰の栽培が始まったのは昭和28 年ごろだといわれています。ぶどうといえば「デラウェア」や「ナイアガラ」が主流だった当時、大粒で味が抜群に良い「巨峰」が登場しました。当初は栽培が難しく安定した生産ができなかったようですが、次第に栽培技術が発達し、安定生産が可能になったことで栽培農家が増え、中野市は日本有数の巨峰の産地となりました。

現在では、5月に出荷する超加温栽培、６～８月に出荷する加温・無加温栽培、９～ 11 月に出荷する露地栽培、11月に出荷する抑制栽培、12 ～ １月に出荷する冷蔵出荷と、5月から翌年１月までの長期間にわたり、味にこだわったぶどうを出荷しています。

仲間との支え合い

地域のぶどう農家には、産地の歴史を築いてきた先輩方だけでなく、若い後継者が入ってきて活気があり、とても心強く感じています。

畑で栽培技術の話をしたり、お互いのハウスのビニールを張る作業をしたりして、仲間同士支え合ってぶどう栽培をしています。昨年のことですが、突風でハウスのビニールが捲れ上がってしまったときに、園主が不在にもかかわらず近所の仲間たちが集まって直してくれたという出来事もありました。

中野のぶどうを全国へ

地元で採れた新鮮でおいしいぶどうを「おつかいもの」などに使っていただき、「信州中野のぶどう」を自慢の逸品として皆さんにも宣伝していただければと思います。

日本一の産地を目指して―。
味にこだわる信州中野のぶどう

武 たけだ田政まさし志
（竹原）
さん

信州中野発！

甘くて果汁がたっぷりのおいしいぶどうはどのように栽培されているのでしょうか？
巨峰の露地栽培の様子を、ＪＡ中野市の営農指導員の方に聞きました。

おいしいぶどうができるまで

②発芽（４月）：春を迎え、ぶどうの木から若い新芽が顔を出します
①剪定（１２月～２月上旬）：枝を切ることにより降雪によるぶどう棚の倒壊を防ぎます

おいしいぶどうは、
農家の皆さんの努力と
愛情の結晶です！

⑩収穫（９月）：手間暇かけて育てたぶどうがついに収穫を迎えます
⑨着色（８月）：巨峰は１粒ずつ着色が進んでいきます
▲ＪＡ中野市でぶどうの営農指導員を務める清しみず水達たつや哉さん

10
別冊
なかの
広報
vol. 1

２．文の配置とテキストの順序がうまく対応しない

- コピー＆ペーストで、文字や文の配置とテキストの順序が不適切になる
  - 次のスライドに定義箇条形式からのテキスト抽出の例をしめす。
- ２段組の２つの段でテキストがつながってしまうことがある
- 表や図のコピー＆ペーストも面倒
  - 表や図はむしろ画像にしてしまう方が良いときもある

例　定義箇条　　Adobe Reader で選択して、コピー

Title	*DITA Open Toolkit User Guide and Reference.*
Edition, release	First edition, August 10, 2006. Based on release 1.2.2 of DITA Open Toolkit.
Publishing information	DITA Open Toolkit is an open source, reference implementation of the OASIS DITA standard (currently DITA 1.0)
Authors	Anna van Raaphorst and Richard H. (Dick) Johnson, principals, VR Communications, Inc. (*http://www.vrcommunications.com*).
Description	This document is the definitive source of information about DITA Open Toolkit (OT). It is also a product of the architecture and the recommended best practices, having been written entirely in DITA XML and produced using the principles and procedures described in the document. With a few minor exceptions (in anticipation of some major enhancements in release 1.3, we made use of the existing bookmap and PDF2 functionality), *DITA Open Toolkit User Guide and Reference* is "vanilla DITA." By processing this document to a given target environment you can see output with no specializations or other special configurations applied. In general, we have found the vanilla outputs adequate for our needs, if not always glamorous or exciting.

※定義と説明文が入り組む　⇩　テキストエディタに貼り付け

Title DITA Open Toolkit User Guide and Reference.
Edition, release First edition, August 10, 2006. Based on release 1.2.2 of DITA Open Toolkit.
DITA Open Toolkit is an open source, reference implementation of the OASIS DITA standard (currently DITA 1.0)
Publishing information
Anna van Raaphorst and Richard H. (Dick) Johnson, principals, VR Communications, Inc. (http://www.vrcommunications.com).
Authors
This document is the definitive source of information about DITA Open Toolkit (OT). It is also a product of the architecture and the recommended best practices, having been
Description
written entirely in DITA XML and produced using the principles and procedures described in the document. With a few minor exceptions (in anticipation of some major enhancements in release 1.3, we made use of the existing bookmap and PDF2 functionality), DITA Open Toolkit User Guide and Reference is "vanilla DITA." By processing this document to a given target environment you can see output with no specializations or other special configurations applied. In general, we have found the vanilla outputs adequate for our needs, if not always glamorous or exciting.

3．文字間の空き、改行の扱い

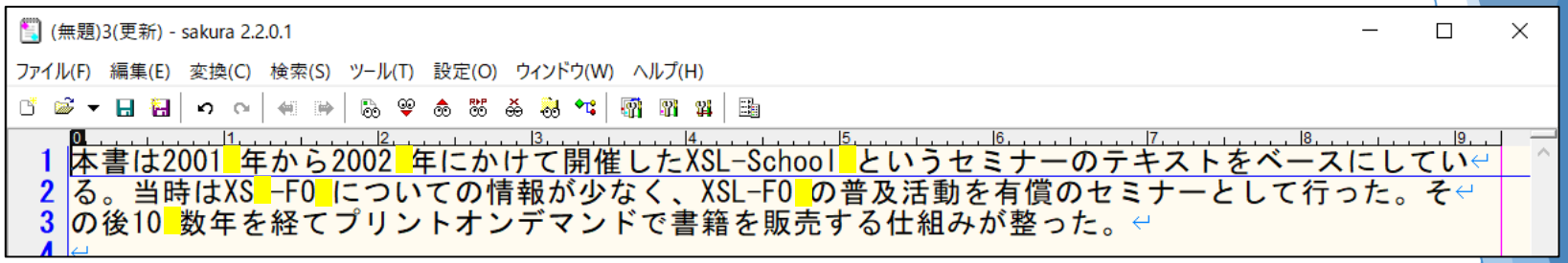
- 和欧混色の場合、元のテキストにはスペースがないが、コピー＆ペースト後にスペースが入る
- 行末の改行の扱い

※コピー＆ペーストを処理するPDFリーダーによる相違がある

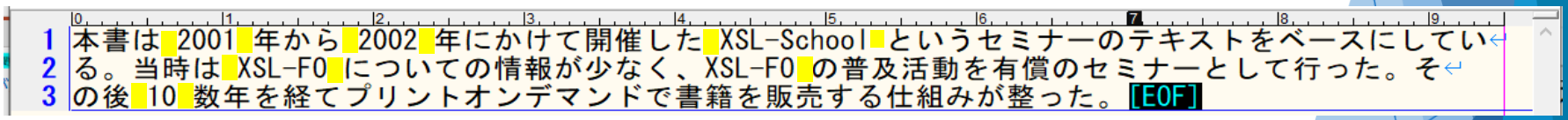
Adobe Reader で選択して、コピー

半角空白が挿入される（黄色のところ）

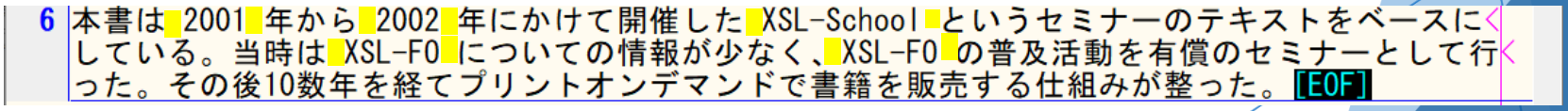
テキストエディタに貼り付け



Google Chromeでコピーすると、半角空白挿入位置が異なる



FireFoxでコピーすると、さらに結果が異なる。行末に改行が入らない。



PDF内のテキストには和文と欧文間の空白文字はない。PDFの表示オペレータ（Tj/TJ）のパラメータで表示位置の調整が行われている。
改行も同様。

４．不要な文字が抽出される

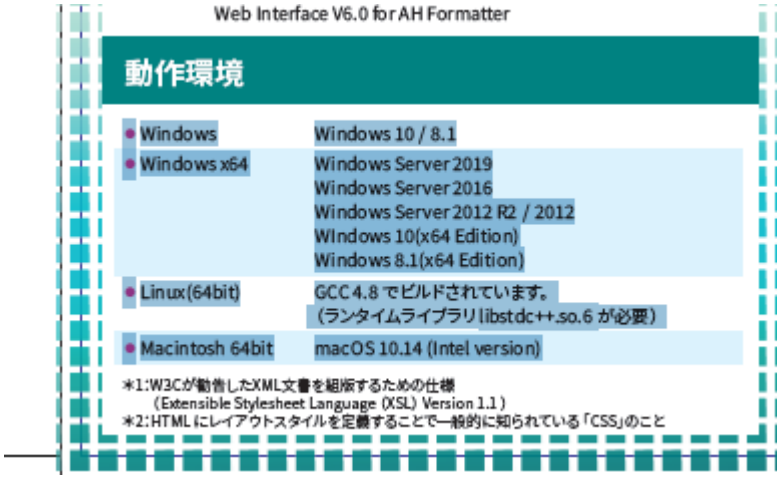
- 文字が2重に出る　⇒　強調を文字の２重表示で表現しているとき、コピー＆ペーストすると文字が2重になる。
  最近のAdobe Readerは余り起きない
- 表示されていない文字が取れる

  ⇒　PDFで非表示に設定されている文字など
- 不必要なテキストがコピーされる
  - ヘッダー・フッター（柱）は抽出不要なので無視したい
  - ページ番号（ノンブル）は抽出不要なので無視したい

5．文字コードを取得できない

※この問題には回避策がないことが多い。

- PDFのテキストを正しい標準文字コードに変換できない
  - カスタムエンコーディング、ユーザー定義文字などが使われている
- グリフがアウトライン化（画像化）されている


Web Interface V6.0 for AH Formatter

動作環境

Windows	Windows 10 / 8.1
Windows x64	Windows Server 2019 Windows Server 2016 Windows Server 2012 R2 / 2012 Windows 10(x64 Edition) Windows 8.1(x64 Edition)
Linux(64bit)	GCC 4.8 でビルドされています。 （ランタイムライブラリ libstdc++.so.6 が必要）
Macintosh 64bit	macOS 10.14 (Intel version)

＊1:W3Cが勧告したXML文書を組版するための仕様
（Extensible Stylesheet Language (XSL) Version 1.1）
＊2:HTMLにレイアウトスタイルを定義することで一般的に知られている「CSS」のこと

↑

Adobe Acrobat Readerで選択、コピー

テキストエディタにペースト⇒

※ラテン文字が抽出できない

でビルドされています。
（ランタイムライブラリ が必要）

コピー＆ペーストには問題多い

- **コピー＆ペーストはPDF内のテキストを再利用するための十分な解決策にならない！**
- その原因は、PDFファイルの作成目的は、テキストを**表示・印刷する**ことにあるため
- 新しいツールが必要！
- **簡単そうで困難な課題**

開発中

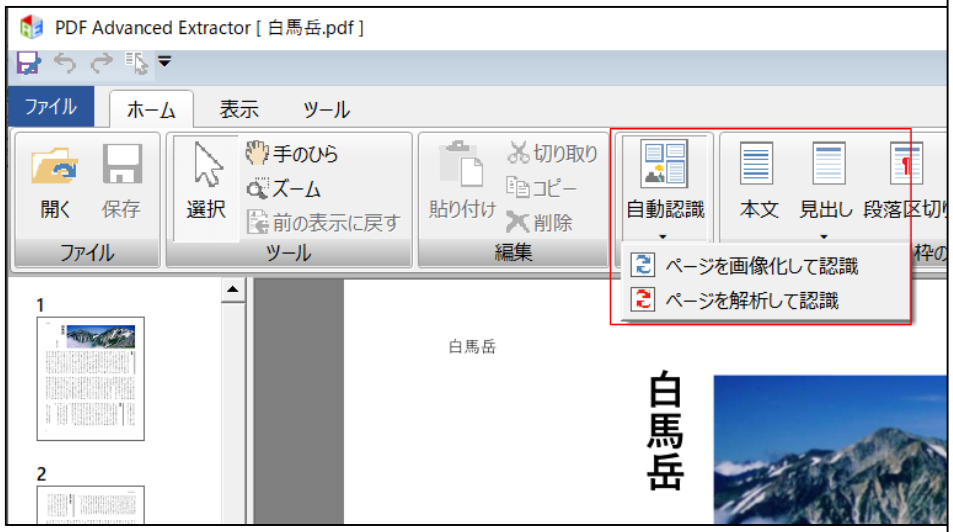
テキスト抽出ツール
PDF Advanced Extractor α版

- PDFファイルからテキストデータを抽出する
- PDFのテキストをコピー＆ペーストで起きる典型的な問題として示した、例１～４については本ツールを使えば解決できる

※　PDFファイルを解析する方法のため、残念ながら、例５は解決できない

テキスト枠を自動認識

- 2段組、段抜きなどの領域を自動認識

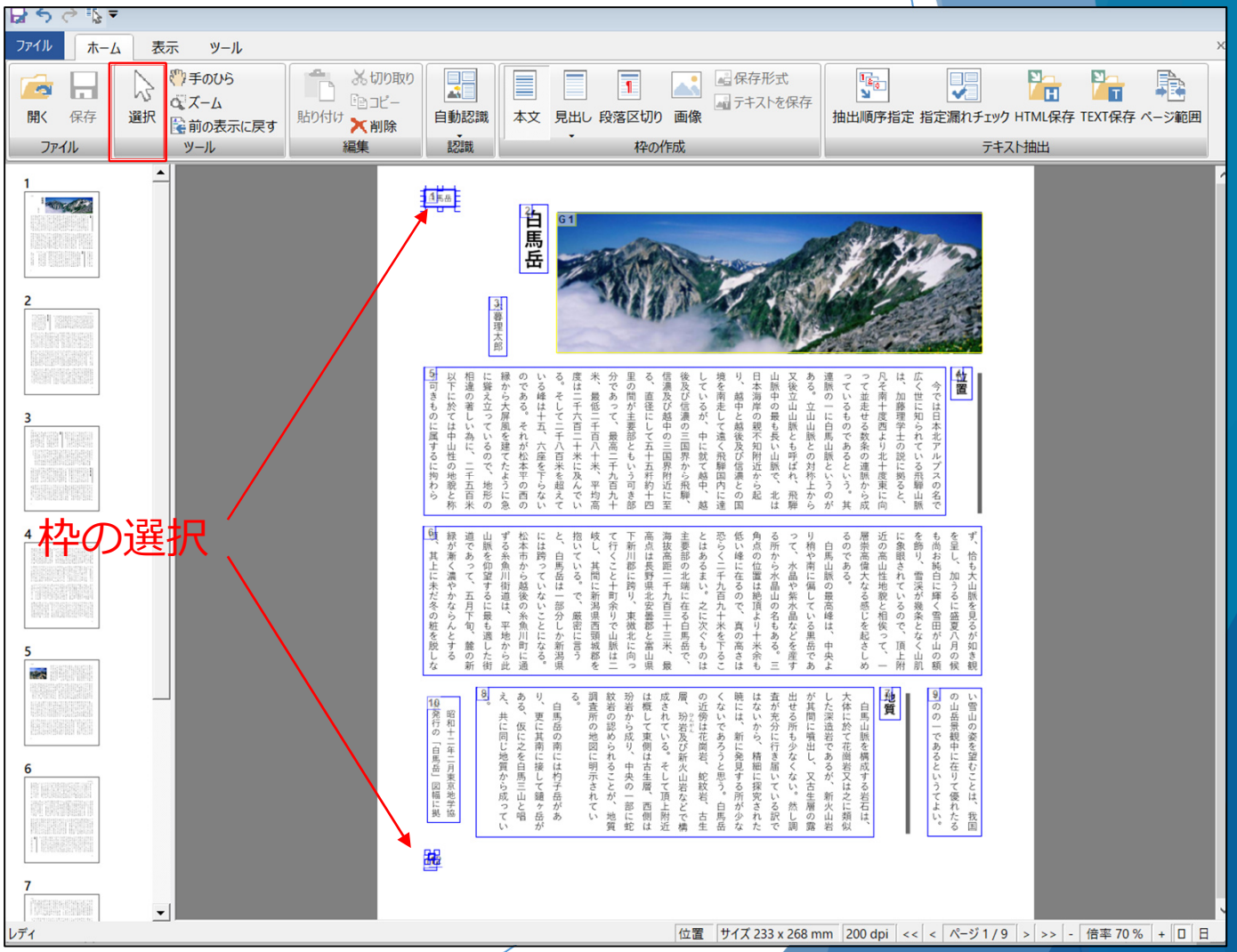


テキスト枠を作成

- 画面上でマウスをドラッグしてテキスト枠を指定
- 見出しの指定もできる

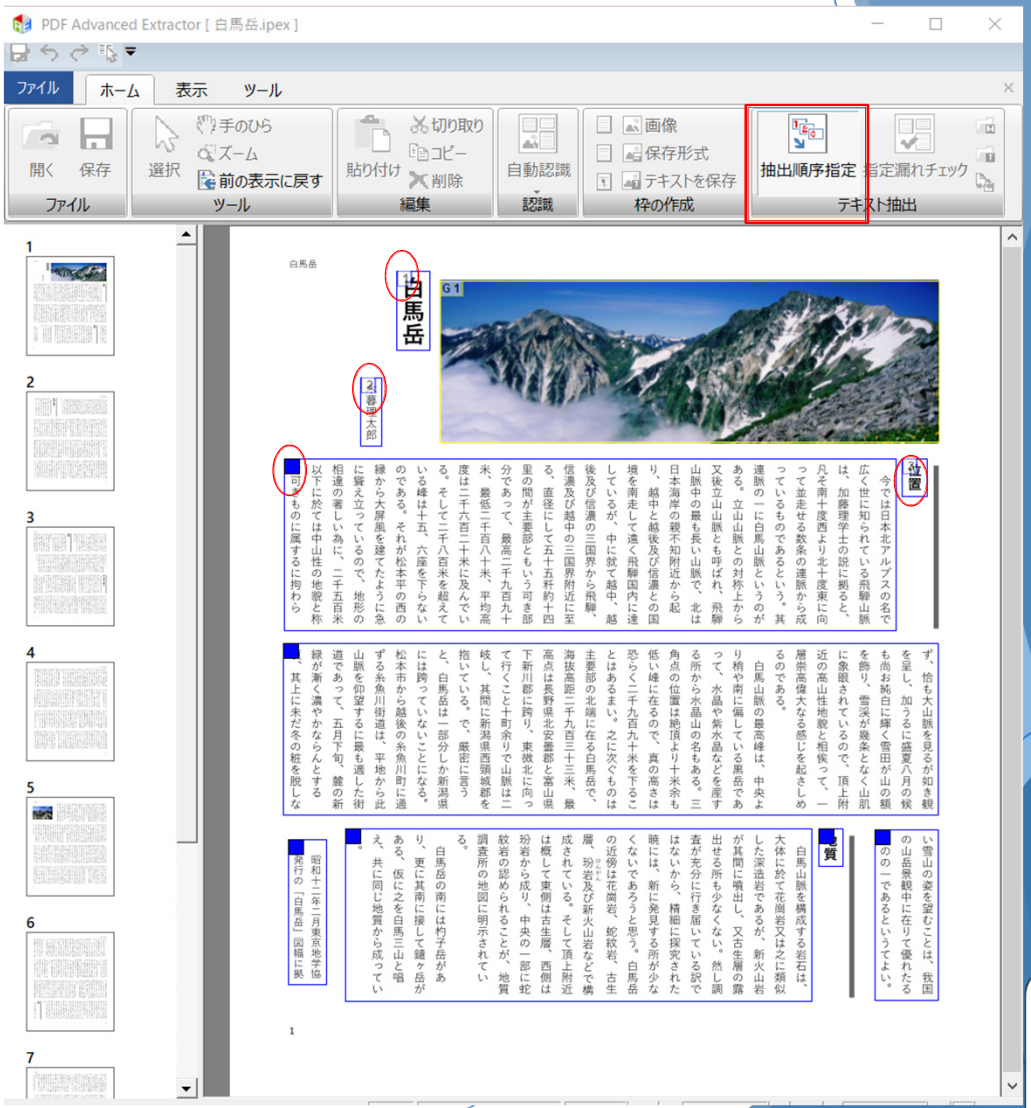
テキスト枠の編集

- 枠選択
- 枠拡大
- 枠縮小
- 枠移動
- 枠結合
- 枠削除



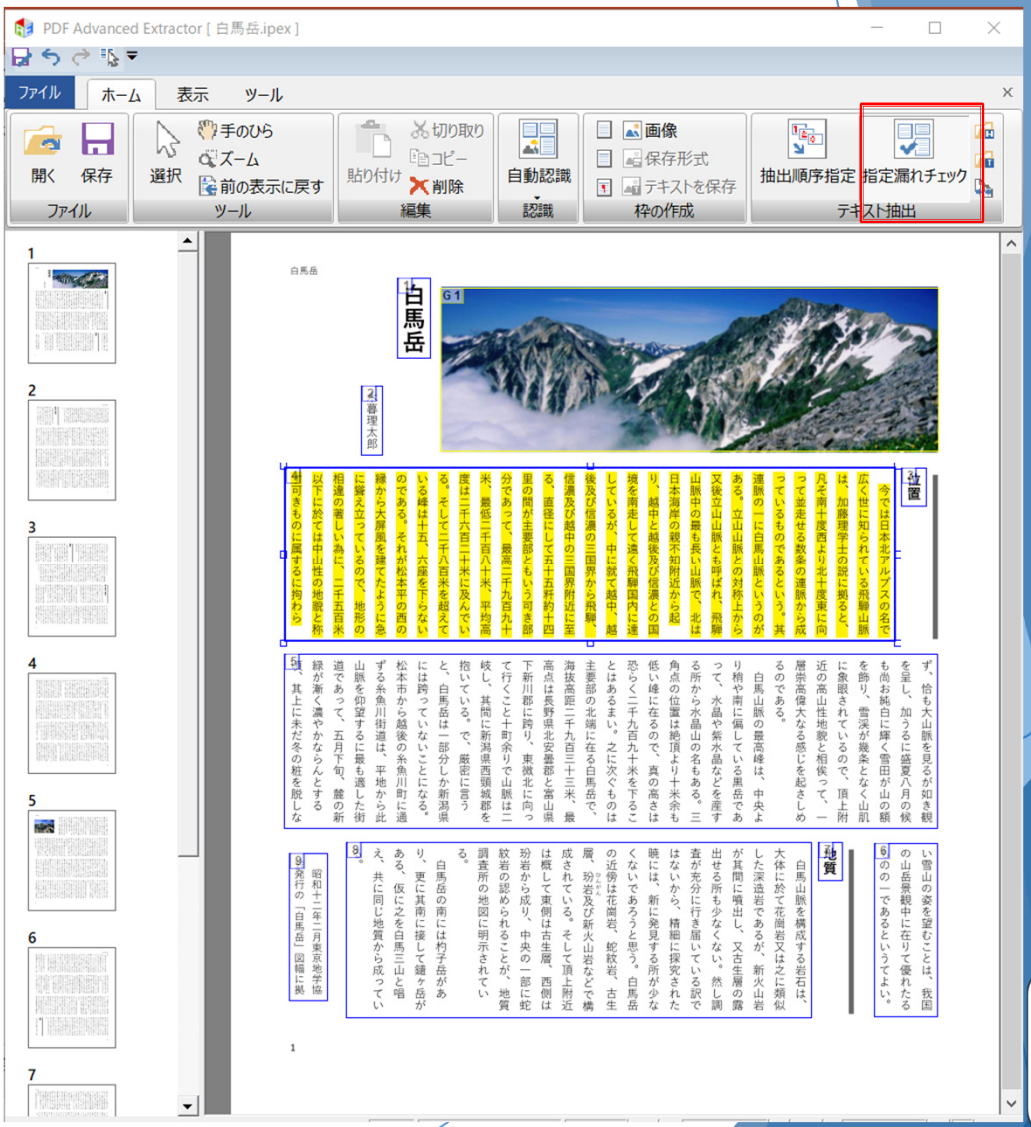
テキスト枠の抽出順序変更

- テキスト枠の内容を出力する順序を指定
- 既定値は作成順。必要に応じて付け替え



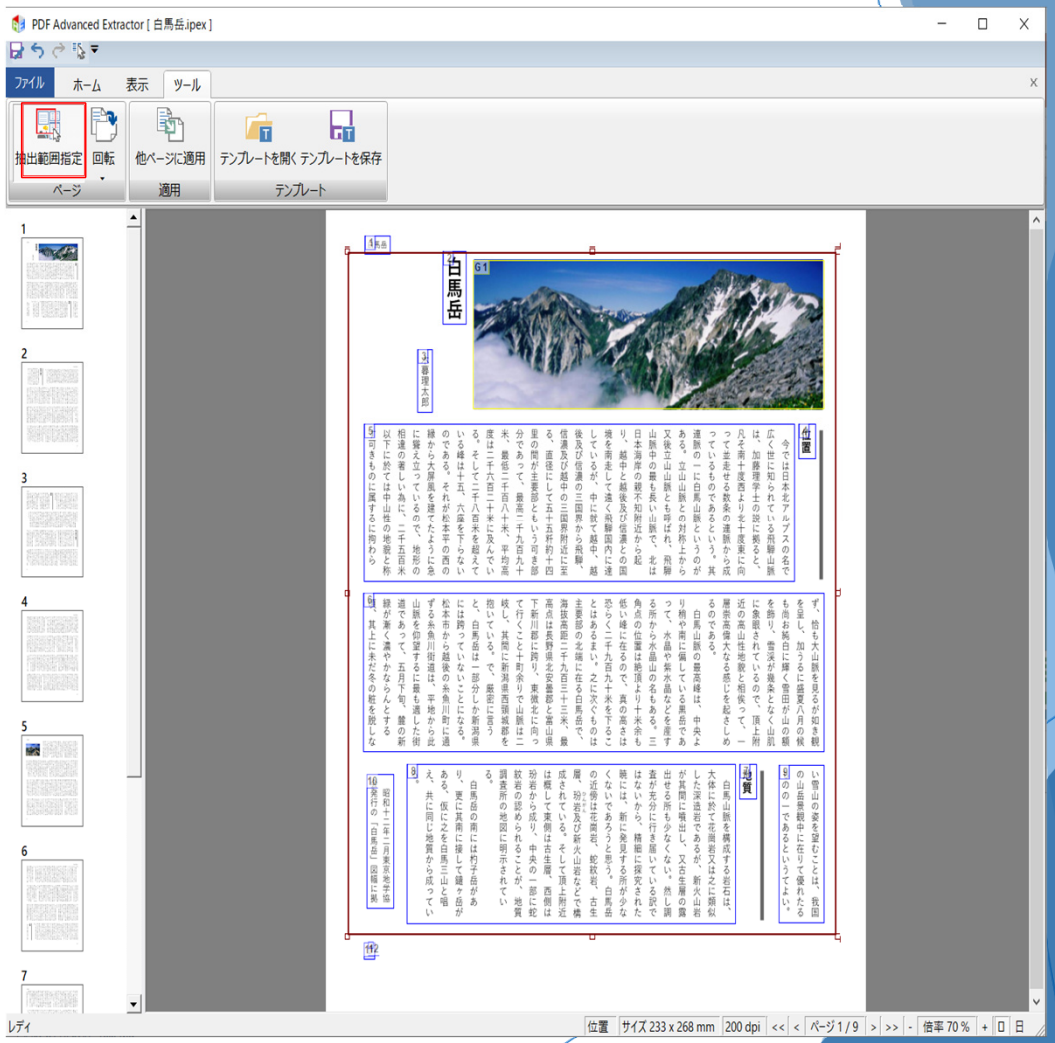
指定漏れチェック

- テキスト枠毎に、文字を抽出したときに脱落しないかを事前に確認できる



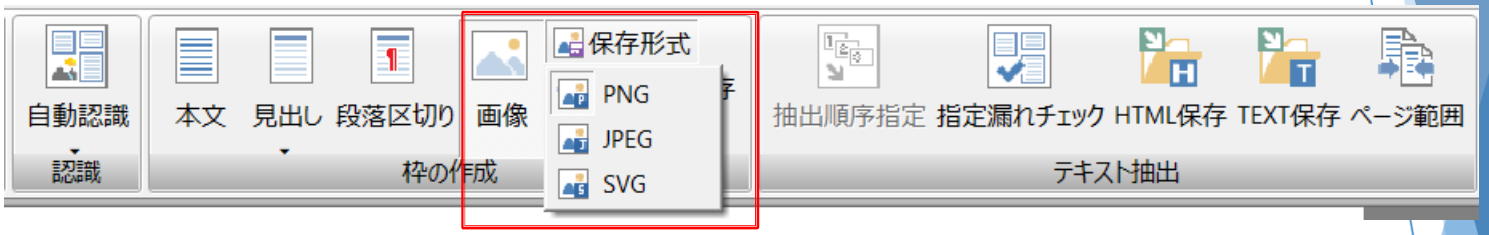
抽出範囲指定

- ページの中でテキストを抽出する範囲を指定する
- 本文領域を範囲指定すると、ヘッダー、フッターを除外できる



画像保存

- 画像領域を指定して画像ファイルを作成



- 選択範囲を画像化。表やグラフなどの画像化に便利

PDF

2-2 サンプル PDF

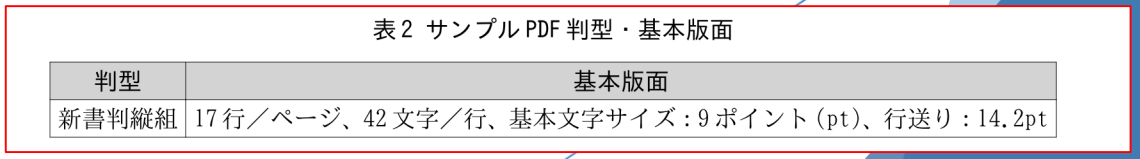
サンプルとして配布している PDF は次のような仕様となっています。

G 1

表 2 サンプル PDF 判型・基本版面

判型	基本版面
新書判縦組	17 行／ページ、42 文字／行、基本文字サイズ：9 ポイント（pt）、行送り：14.2pt

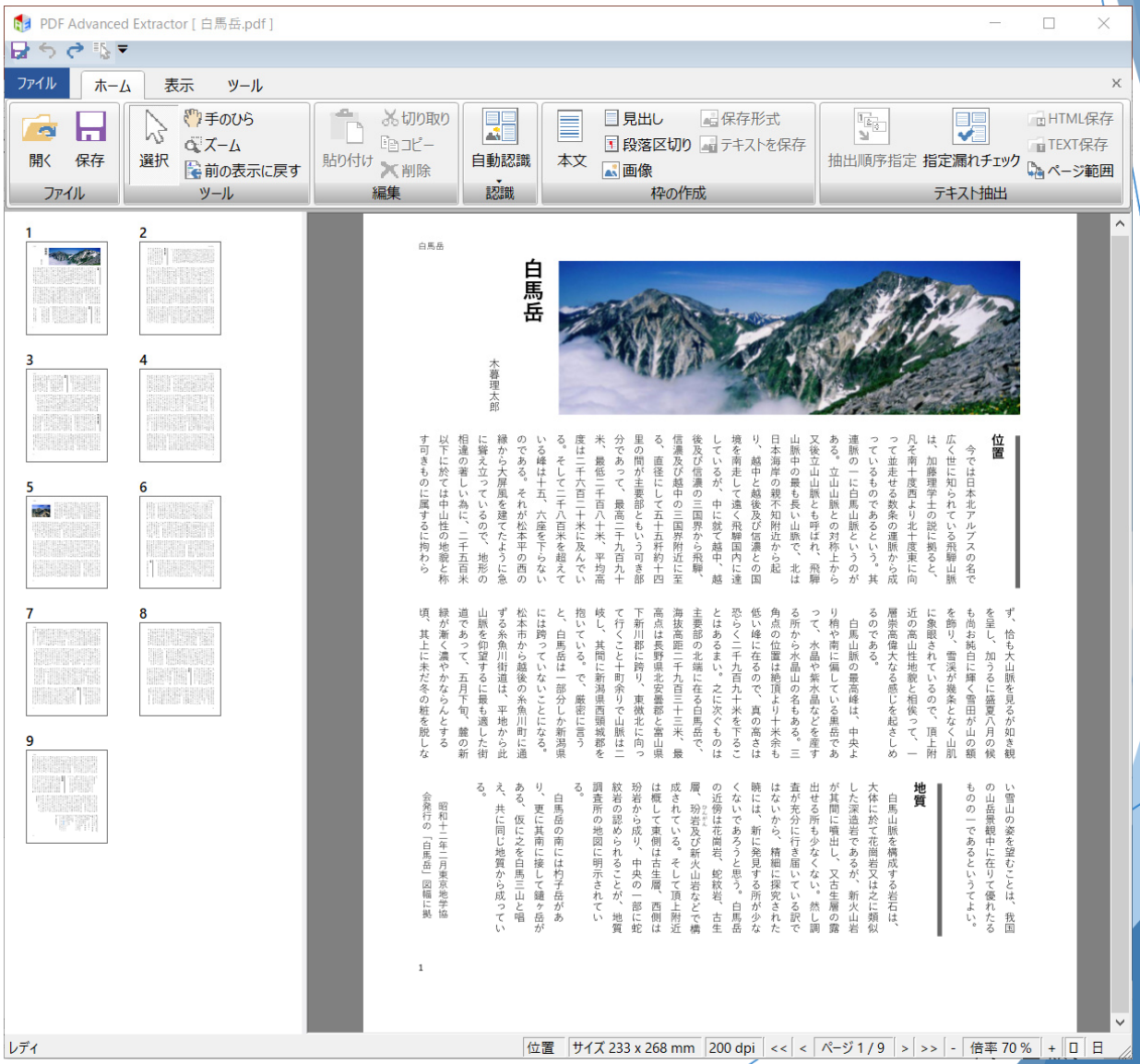
PNG



HTMLタグ付き保存

- HTMLタグ付きテキストでの保存
- bodyの下位に出力するタグは次の8種類
  - 段落タグ（p）
  - 見出しタグ（h1～h6）
  - イメージタグ（img）
  - イメージ範囲にテキストがあるとき、テキストをimgタグのalt属性値に保存できる

デモ


PDF Advanced Extractor [白馬岳.pdf]
ファイル
ホーム
表示
ツール
開く
保存
選択
手のひら
ズーム
前の表示に戻す
貼り付け
切り取り
コピー
削除
自動認識
本文
見出し
段落区切り
画像
保存形式
テキストを保存
抽出順序指定
指定漏れチェック
HTML保存
TEXT保存
ページ範囲
ファイル
ツール
編集
認識
枠の作成
テキスト抽出
白馬岳
木暮理太郎
位置
地質
レディ
位置
サイズ 233 x 268 mm
200 dpi
ページ 1 / 9
倍率 70 %

製品概要

項目	内容
製品名	PDF Advanced Extractor
バージョン	α版（仕様が変更される可能性があります）
動作環境	Windows 10　64ビット版
種別	デスクトップ製品　（サーバー上で使ったり、システムに組み込むことはできません）
価格	オープン価格
発売日	未定
製品Web	https://www.antenna.co.jp/pdfae/

α版試用と評価のお願い

- 本製品のα版をお試しください。
- 評価結果をお待ちしています。

次のURLから評価版をダウンロードいただけます：

https://www.antenna.co.jp/pdfae/eval.html

お問い合わせ先：oem@antenna.co.jp

参考資料

アンテナハウスPDF資料室

- PDFからテキストを取り出す処理について
  「プログラマーから見たPDFファイル」
  https://www.antenna.co.jp/pdf/reference/pdftext.html
- PDFテキスト抽出の諸問題
  「簡単そうで簡単ではないPDFのテキスト抽出」
  https://www.antenna.co.jp/pdf/reference/text-extractor.html