MVC 推進部様向け

OCR 業務システム 操作説明書

Ver 1.0.0

作成者：しもベッチ

作成日：2026 年 6 月 30 日

目次

1. 本書について 5
1.1 はじめて使う方へ 5
2. システムへのアクセスとログイン 6
2.1 システムへのアクセス 6
2.2 新規登録とログイン 6
2.3 初回ログイン時のプロフィール登録 8
3. 画面共通の操作 11
3.1 共通ヘッダー 11
4. 基本操作の流れ 12
5. OCR 結果画面の確認 15
5.1 抽出されたテキスト 17
5.2 調査用の折りたたみ情報 18
5.3 行分割データの見方 22
5.4 整形前テキストの見方 22
5.5 OCR 結果画面から次に進む判断 22
6. 帳票 config 編集と OCR 再実行 23
6.1 config 編集後の再 OCR の流れ 29
6.2 config 編集画面の主な項目 29
6.2.1 Simple（テキスト）レイアウトの調整方法 32
6.2.2 表組（TypeA / TypeB）レイアウトの調整方法 34
6.3 AI による OCR 文字揺れ辞書作成 38
7. Simple（テキスト） / 表組（TypeA） / 表組（TypeB） の使い分け 41
7.1 Simple（テキスト）レイアウト 41
7.2 表組（TypeA） / 表組（TypeB） 42
7.3 運用上の使い分け 42
8. キーワード解析結果画面の確認 43
8.1 Major / Minor / Value の考え方 46
8.2 対応する Excel 項目名の確認 47
8.3 丸数字付き仮受け項目 47
8.4 Excel 出力時の特殊処理 47
8.5 Value の確認と修正 47
8.6 手修正と再 OCR の使い分け 48

9. Excel / CSV 出力 49
9.1 出力ファイル名 49
10. 管理画面の基本操作 50
10.1 管理画面ダッシュボードと権限 50
10.2 左メニュー 50
10.3 一覧画面の共通操作 51
10.4 編集画面 53
10.5 管理画面右上のユーザーメニュー 58
11. 困ったときの確認ポイント 59
Appendix A. OCR パーサーの種類と表構造について 60
A.1 OCR パーサーの役割 60
A.2 表の配置範囲 60
A.3 Simple（テキスト）レイアウト 61
Simple（テキスト）レイアウトの考え方 61
Simple（テキスト）レイアウトが向いている帳票 62
Simple（テキスト）レイアウトが向いていない帳票 62
Simple（テキスト）レイアウトで調整できること 62
備考ブロックでの注意 63
A.4 表組（TypeA）レイアウト 63
A.5 表組（TypeB）レイアウト 63
A.6 赤線・青線の指定方法 64
A.7 レイアウト選択の目安 64
Appendix B. OCR パーサーが出力する正規化データについて 65
B.1 正規化データの目的 65
B.2 正規化された行出力の基本形式 65
B.3 行タイプと省略形 65
B.4 低信頼マーク「？」の扱い 66
B.5 major / minor / value の意味 66
B.6 備考・備考モードの扱い 66
B.7 ヘッダ・フッタの扱い 66
B.8 Simple（テキスト）レイアウトの場合 67
B.9 正規化データを見るときの確認ポイント 67
Appendix C. 帳票 config パラメーター一覧 71
C.1 調整の基本 71

C.2 パラメーター一覧 72
Appendix D. OCR 揺れに関する変換辞書について 76
D.1 translate_dic（帳票別カスタム辞書）の位置づけ 76
D.2 word_variant_map の基本登録ルール 76
D.3 登録してよい語 76
D.4 登録してはいけない語 77
D.5 反復変換で安全な定義 77
D.6 長文辞書を残す条件 77
D.7 実務上の判断基準 77
D.8 構造付き辞書の登録ルール 78
D.9 構造付き辞書の例と注意点 78
D.10 辞書追加後の確認 79
Appendix E. CSV 出力時の項目展開と特殊項目について 80
E.1 Excel 項目マスターと出力項目 80
E.2 出力時の扱いの分類 80
E.3 未設定の扱い 80
E.4 重複項目の扱い 80
E.5 丸数字付き仮受け項目 80
E.6 住所系の分解 81
E.7 年月系の分解 81
E.8 インポート対象とインポート外項目 81
E.9 出力前の確認ポイント 81
改訂履歴 82

1. 本書について

本書は、MVC 推進部 OCR 業務システムの操作方法を説明するものです。PDF 帳票をアップロードし、OCR 解析、抽出結果確認、キーワード解析、Excel / CSV 出力までを行うための基本操作を記載します。

また、帳票ごとの読み取り結果を調整するための帳票 config 編集画面、Simple / 表組（TypeA） / 表組（TypeB） の使い分け、管理画面の基本操作についても説明します。

項目	内容
システム名	MVC 推進部 OCR 業務システム
対象帳票	不動産関連 PDF 帳票。主な区分は建物全部、建物一部、土地。
主な出力	基幹システム連携用の Excel / CSV 出力。
対象者	一般利用者、帳票設定担当者、システム管理者。

本書の読み方

通常の利用者は、ログイン、PDF アップロード、OCR 結果確認、キーワード解析、Excel / CSV 出力の章を確認してください。

帳票設定を調整する担当者は、帳票 config 編集、Simple / 表組（TypeA） / 表組（TypeB）、管理画面の章も確認してください。

1.1 はじめて使う方へ

本書は、通常の OCR 操作から帳票設定・管理画面の操作までをまとめています。すべての章を最初から順番に読む必要はありません。利用目的に応じて、先に確認する章を選んでください。

利用目的	先に確認する章	内容
通常の Excel / CSV 出力だけを行う場合	2、3、4、5、8、9	ログイン、PDF アップロード、OCR 結果確認、キーワード解析、Excel / CSV 出力までの基本操作です。
OCR 結果が崩れる場合、新しい帳票を調整する場合	6、7、11	帳票 config 編集、再 OCR、Simple / 表組（TypeA） / 表組（TypeB）の使い分け、困ったときの確認方法を説明しています。
管理画面やマスター情報を確認・管理する場合	10	管理画面の一覧、編集、CSV 取込・出力、レコード複写などの基本操作を説明しています。
まずは通常操作の章だけを確認し、読み取り結果が崩れた場合や帳票設定を変更する場合に、config 調整や管理画面の章を参照してください。		

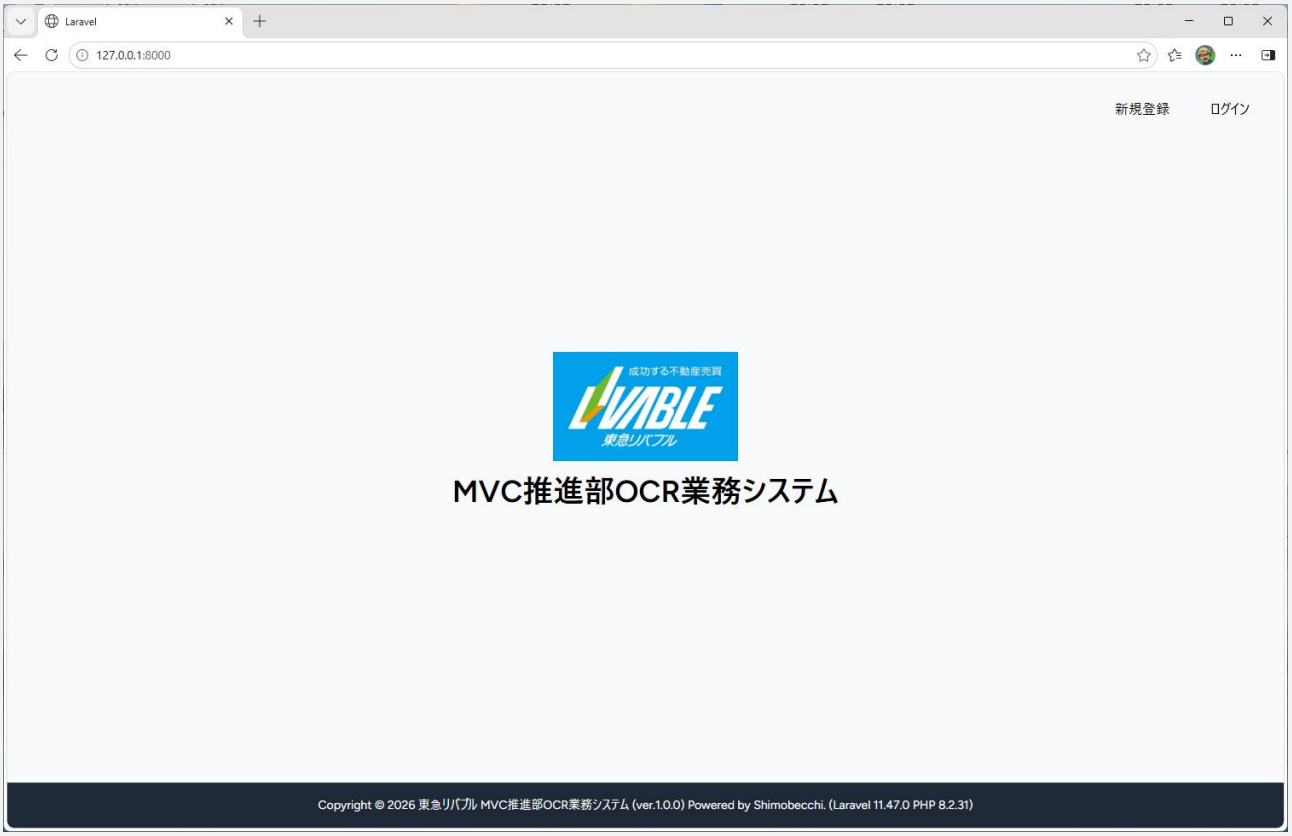
2. システムへのアクセスとログイン

2.1 システムへのアクセス

ブラウザを起動し、以下の URL へアクセスします。

https://TLIAROCRSV11

本システムは社内ネットワーク専用のシステムです。



図：ウェルカム画面

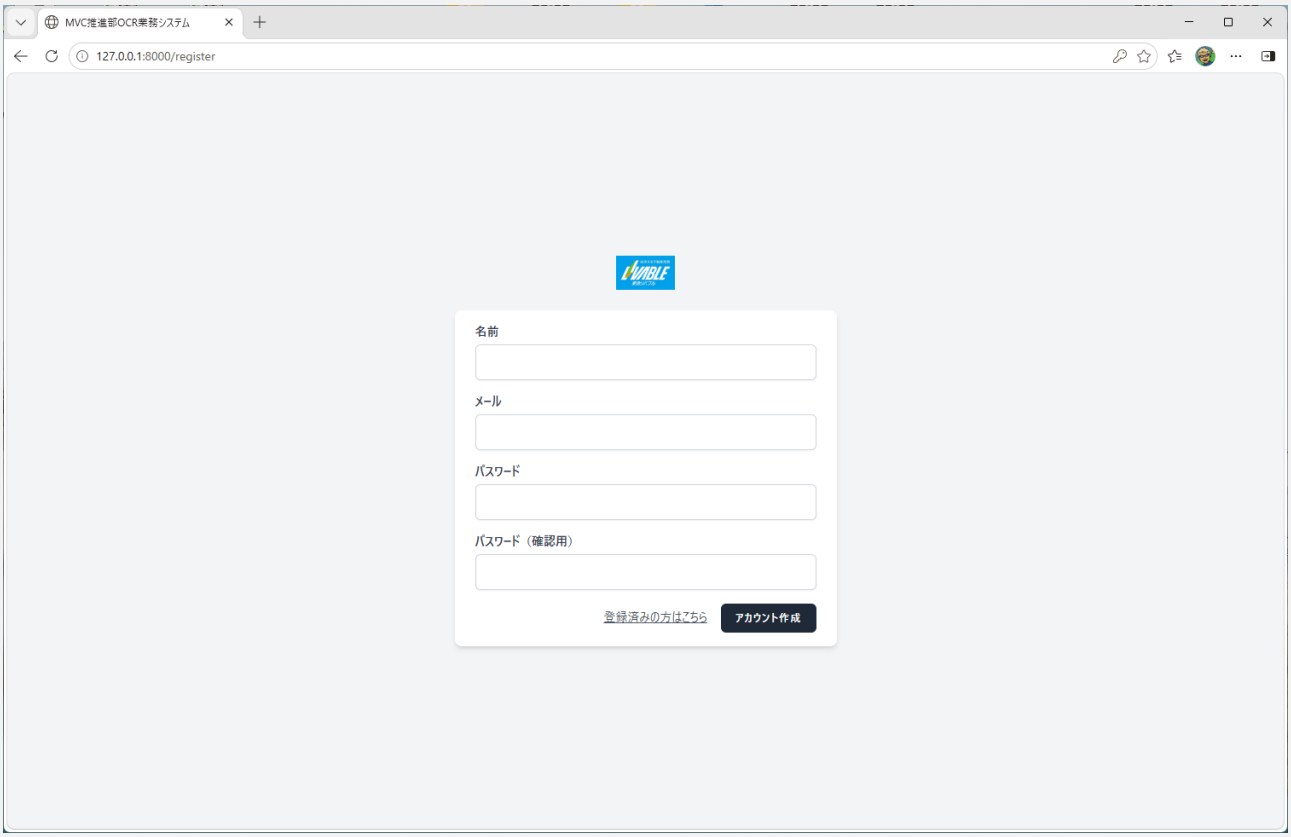
2.2 新規登録とログイン

システムにアクセスすると、最初にウェルカム画面が表示されます。画面右上には「新規登録」と「ログイン」のリンクがあります。

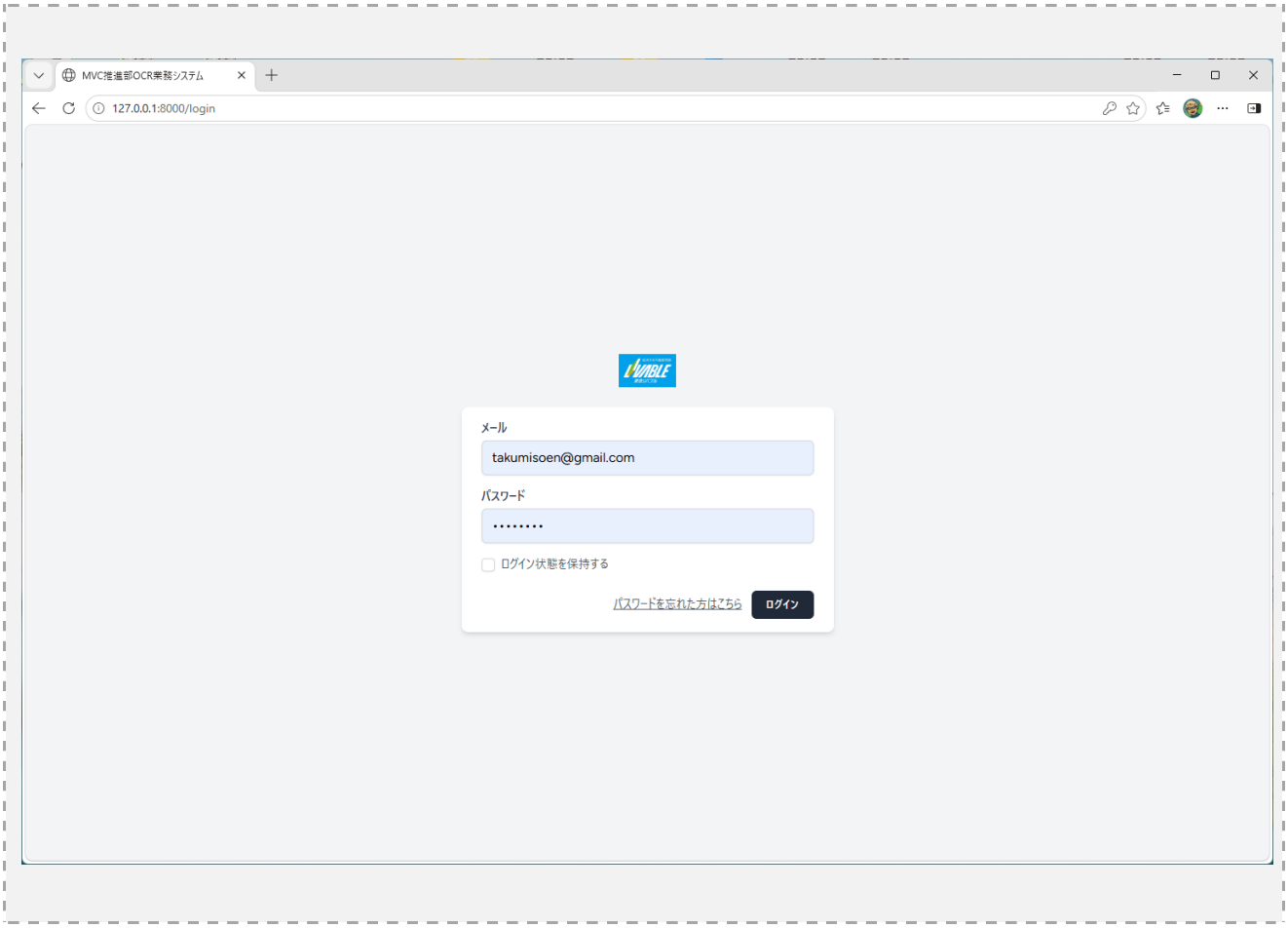
- まだアカウントを持っていない場合は、「新規登録」をクリックします。
- すでにアカウントを持っている場合は、「ログイン」をクリックします。

画面	操作内容
新規登録画面	名前、メールアドレス、パスワード、パスワード確認を入力し、「アカウント作成」をクリックします。
ログイン画面	メールアドレスとパスワードを入力し、「ログイン」をクリックします。必要

画面	操作内容
	に応じて「ログイン状態を保持する」にチェックを入れます。
パスワード再設定	パスワードを忘れた場合は、「パスワードを忘れた方はこちら」から再設定します。



図：新規登録画面



図：ログイン画面

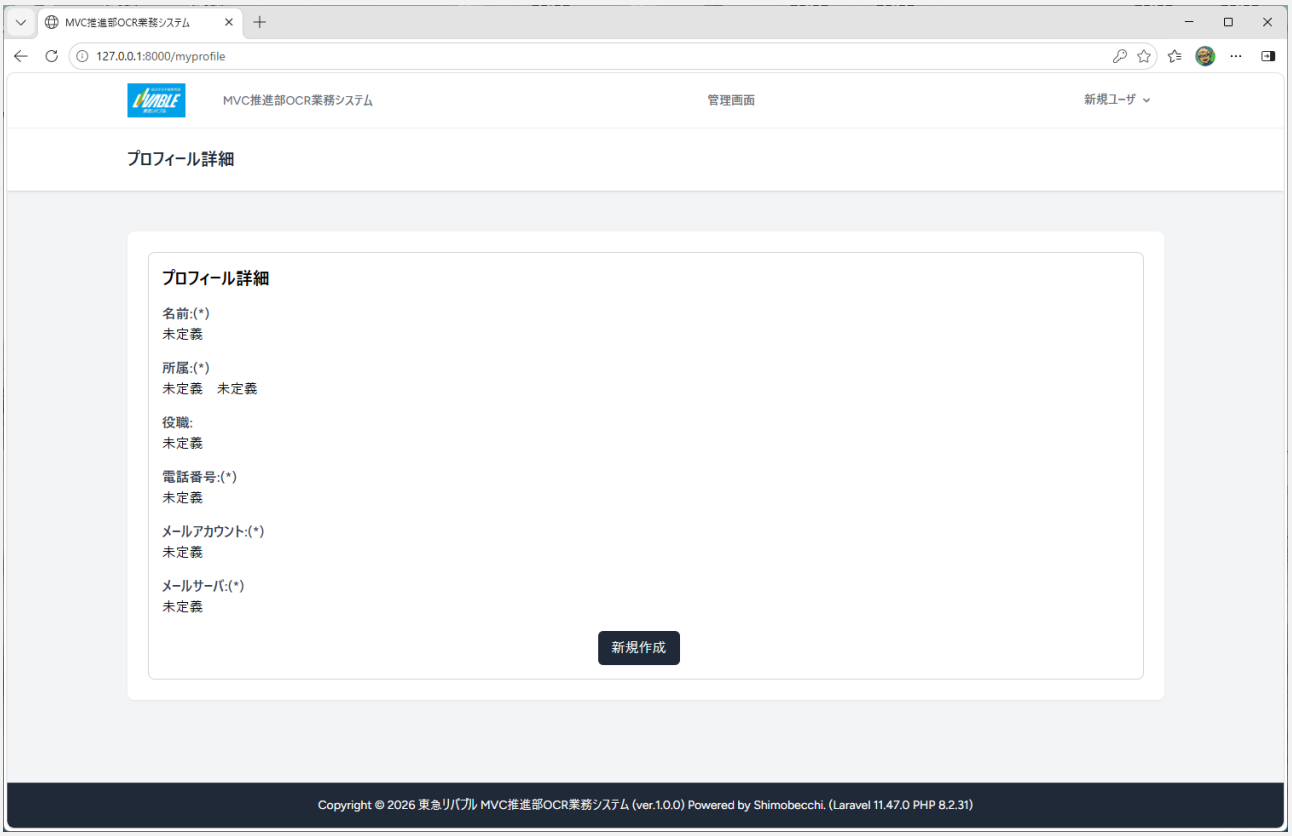
ログインが完了すると、ダッシュボード画面へ移動します。一度ログインした状態では、システムへアクセスしたときにウェルカム画面ではなく、ダッシュボード画面が表示されます。

2.3 初回ログイン時のプロフィール登録

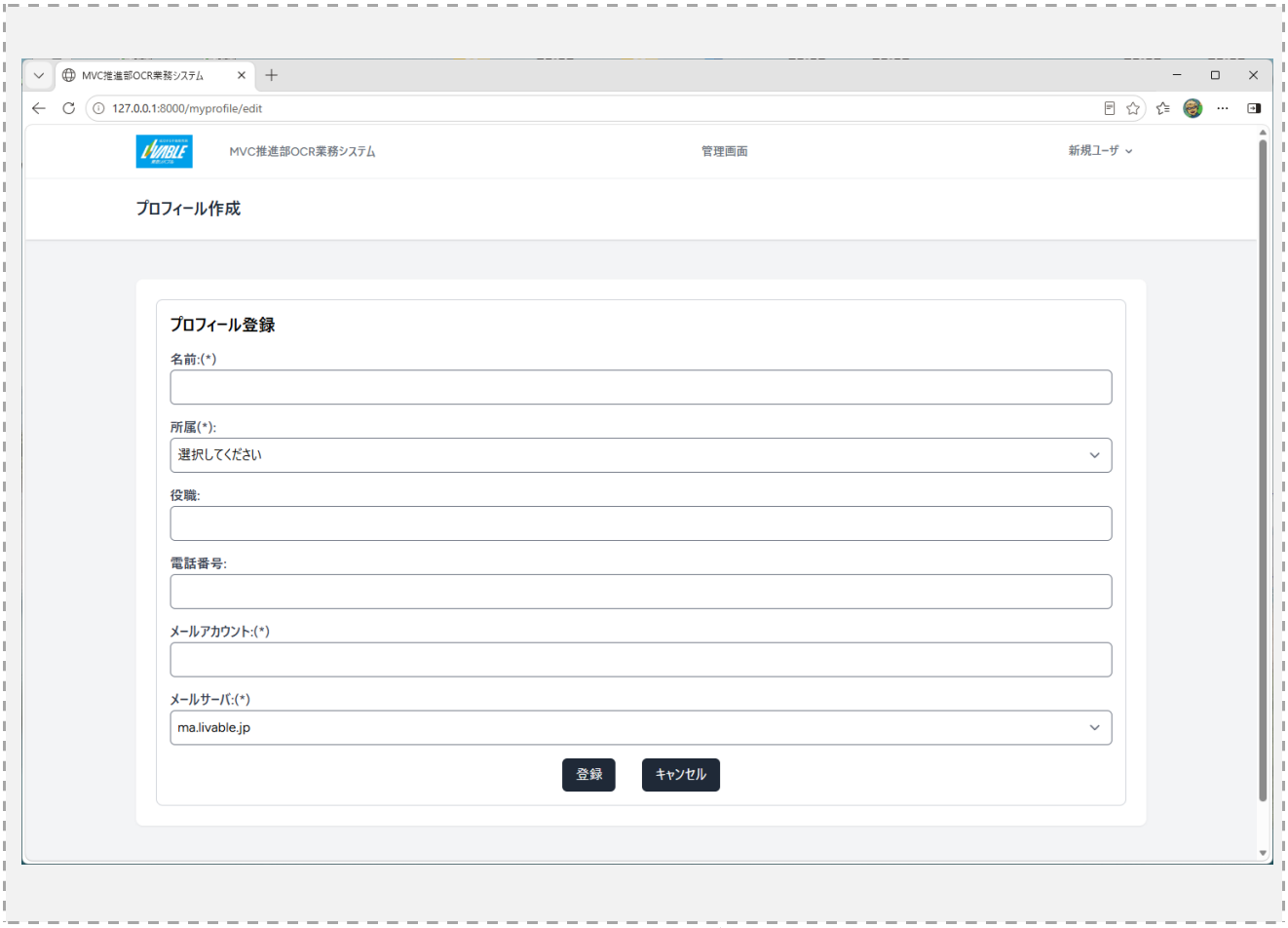
プロフィール情報が未登録のユーザーがログインした場合、ダッシュボード画面ではなく、プロフィール作成画面が表示されます。

> **注意**
> プロフィールを作成するまで、ダッシュボード画面へ進むことはできません。

プロフィール作成画面では、名前、所属、役職、電話番号、メールアカウント、メールサーバーなど、本システム固有の利用者情報を入力します。必要な項目を入力し、「作成」ボタンをクリックしてください。



図：プロフィール未登録時のプロフィール作成画面



図：プロフィール詳細画面

登録済みのプロフィールを変更する場合は、画面右上のユーザー名をクリックし、「プロフィール」を選択します。プロフィール詳細画面で「編集」をクリックすると、登録内容を変更できます。

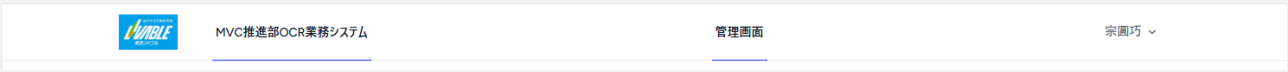
3. 画面共通の操作

3.1 共通ヘッダー

各画面上部には、共通のヘッダーが表示されます。左側の東急リバブルのロゴ、または「MVC 推進部 OCR 業務システム」をクリックすると、ダッシュボード画面へ戻ることができます。

中央付近の「管理画面」をクリックすると、管理画面へ移動します。管理画面では、帳票設定や各種マスター情報の確認・管理を行います。管理画面の利用範囲は、ユーザーの権限によって異なります。

右側にはログイン中のユーザー名が表示されます。ユーザー名をクリックすると、アカウント、プロフィール、ログアウトのメニューが表示されます。



図：共通ヘッダーとユーザーメニュー

メニュー	説明
アカウント	ログイン用アカウント情報やパスワードを管理します。仮パスワードで案内された場合は、初回ログイン後に自分のパスワードへ変更してください。
プロフィール	所属、役職、電話番号、メールアカウント、メールサーバーなど、本システム固有の利用者情報を確認・登録します。
ログアウト	作業を終了し、システムからログアウトします。

4. 基本操作の流れ

本システムの基本的な流れは、PDF アップロード、OCR 実行、OCR 結果確認、キーワード解析、Excel / CSV 出力です。

順序	操作	説明
1	PDF アップロード	対象 PDF を選択してアップロードします。
2	OCR 実行	PDF ページを OCR 解析します。
3	OCR 結果確認	抽出されたテキスト、行分割、config 情報などを確認します。
4	config 調整	必要に応じて帳票設定を修正し、OCR を再実行します。
5	キーワード解析	OCR 結果を基幹システム用の Excel 項目へマッピングします。
6	Excel / CSV 出力	確認済みの内容を CSV として出力します。

新規帳票の標準的な確認手順

新規帳票や未調整帳票では、最初から OCR 結果が正しく出るとは限りません。

表組み 表組（TypeA） / 表組（TypeB） では、赤線・青線の設定が必要になるため、通常は初回 OCR、config 調整、再 OCR の最低 2 回の OCR 実行が必要になります。

Simple（テキスト）レイアウトでは赤線・青線を使用しないため、帳票によっては 1 回の OCR 実行で利用できる場合があります。



図：ダッシュボード画面



図：PDF アップロード画面

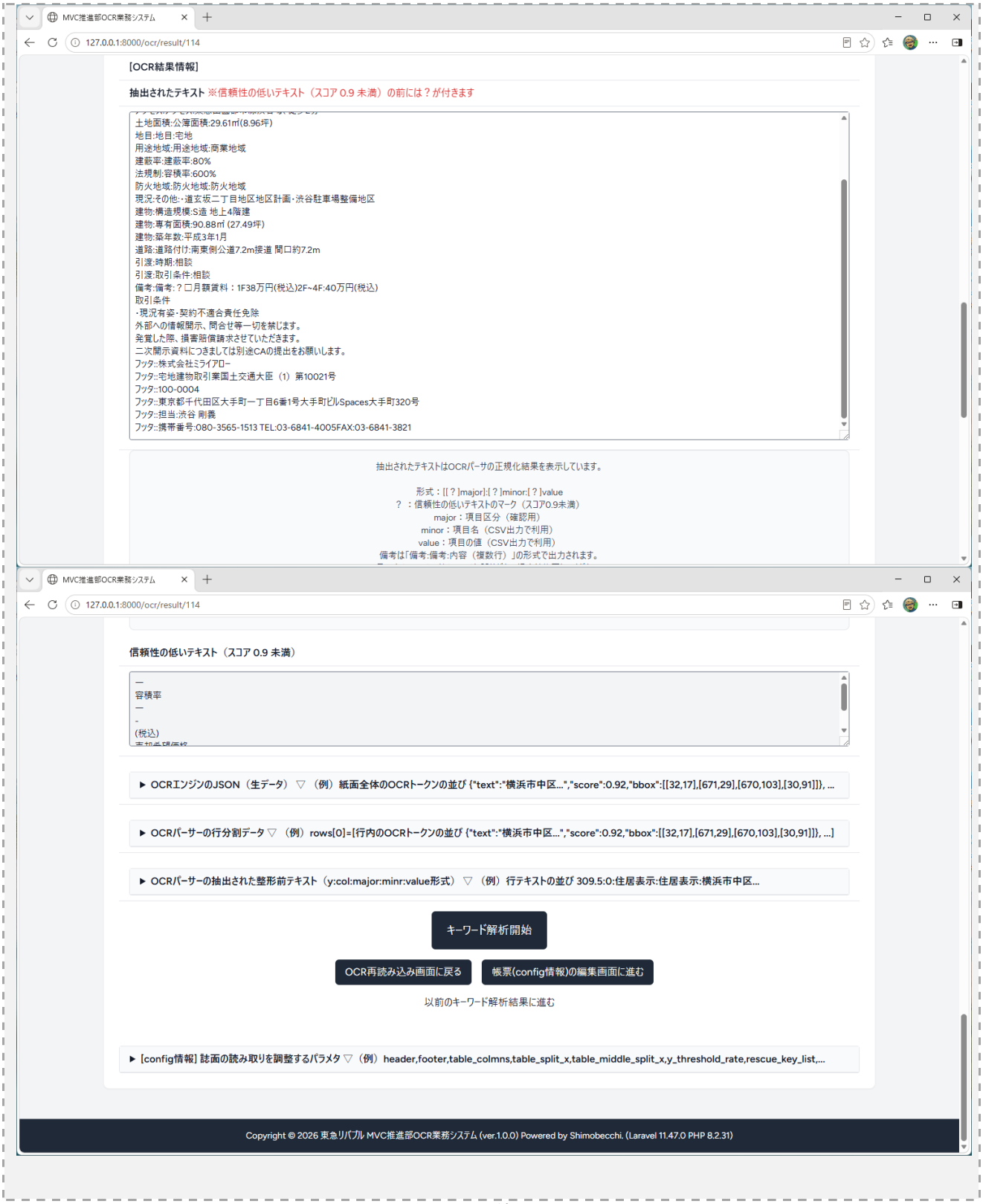


図：アップロード結果画面

5. OCR 結果画面の確認

OCR 結果画面では、読み取った紙面画像、抽出されたテキスト、低信頼マーク、調査用情報、次に進むためのボタンを確認します。


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/114
MVC推進部OCR業務システム
管理画面
宗園巧
MVC推進部OCR業務システム
不動産・マンション情報の自動抽出ファイル（1棟収益.pdf）P.1 の 物件概要書(ミライアロー) のOCR結果
不動産・マンション情報の自動抽出ファイルのOCR結果
x=--- y=---
table_colmns=1
click: x=--- y=---
物 件 概 要 書
物件名称
ミノハナビル
所在地
住所
東京都渋谷区道玄坂2丁目16-1
地番
東京都渋谷区道玄坂2-152-6
アクセス
JR山手線
渋谷
駅
徒歩 3 分
東急田園都市線
渋谷
駅
徒歩 2 分
実測面積
-
地目
宅地
用途地域
商業地域
[PDFページ情報]
ファイル名
1棟収益.pdf
区分（帳票）
建物全部
帳票名
物件概要書(ミライアロー)
OCRレイアウトパーサー
table_layout_typeB
PDFページ数
58
SCANページ番号
1
SCANページ数
1
SCANページサイズ(mm)
74.09 × 104.78
SCANページ向き
P（ポートレイト）
SCANページ回転
-
SCAN画像の表示
表示リンク（1棟収益.pdfの1ページ目の画像）
SCAN画像の幅(px)
1273
SCAN画像の高さ(px)
1800
SCAN画像の長辺サイズ(px)
1800



図：OCR 結果画面

5.1 抽出されたテキスト

抽出されたテキストは、原則として次の形式で表示されます。

[？]major:[？]minor:[？]value

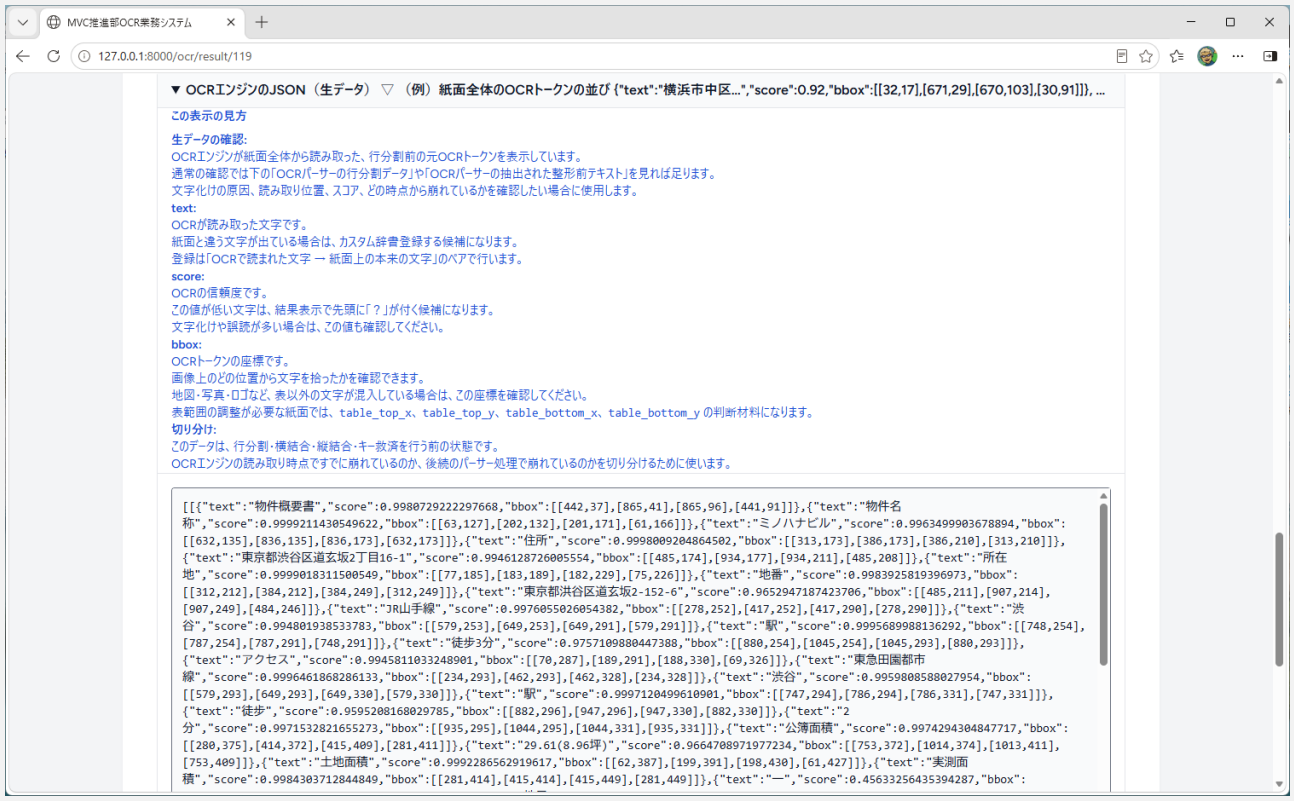
要素	意味
major	大項目。項目の分類や周辺情報を示す参考情報です。
minor	項目名。CSV 出力で特に重要な項目です。
value	項目の値。CSV 出力される内容です。
？	低信頼マーク。OCR 信頼度が低い文字に付く確認用マークです。

キーワード解析時には、低信頼マーク「？」は自動的に除去されます。値に誤りがある場合は、後続のキーワード解析結果画面で Value を修正できます。

5.2 調査用の折りたたみ情報

OCR 結果画面には、通常の操作では毎回確認する必要のない調査用情報が用意されています。これらは、抽出結果に不自然な点がある場合に、どの段階で問題が発生しているかを確認するための情報です。

情報	確認内容
OCR エンジンの JSON（生データ）	OCR エンジンが最初に読んだ元データを確認します。text、score、bbox などを確認できます。
OCR パーサーの行分割データ	OCR トークンが正しい行単位に分かれているかを確認します。
OCR パーサーの抽出された整形前テキスト	major / minor / value に整形される前の中間データを確認します。
config 情報	帳票ごとの設定値と調整方法を確認します。



図：OCR エンジンの JSON（生データ）



図：OCR パーサーの行分割データ



図：OCR パーサーの抽出された整形前テキスト

5.3 行分割データの見方

行分割データでは、OCR で読み取ったトークンが rows[n] 単位でどのように行分割されたかを確認できます。抽出結果で行のまとまりがおかしい場合に確認します。

現象	考えられる原因	主な調整
本来 1 行の文字列が複数行に分かれている	行分割が細かすぎる	y_threshold_rate を大きくします。
本来別行の文字列が同じ行に巻き込まれている	行分割の範囲が広すぎる	y_threshold_rate を小さくします。
同じトークンが複数行に重複して出ている	行間判定が広すぎる	y_threshold_rate を小さくします。
y_threshold_rate で改善しない	行判定基準が合っていない	y_threshold_row_anchor_mode を変更して再調整します。

5.4 整形前テキストの見方

整形前テキストでは、OCR 結果が最終的な抽出テキストに整形される前の中間データを確認できます。内部的には、次の形式で扱われます。

y:col:major:minor:value

y は行位置、col は列番号、major は大項目、minor は項目名、value は値です。この情報は、横結合、縦結合、キー救済、辞書変換などの処理が正しく行われているかを確認するために使用します。

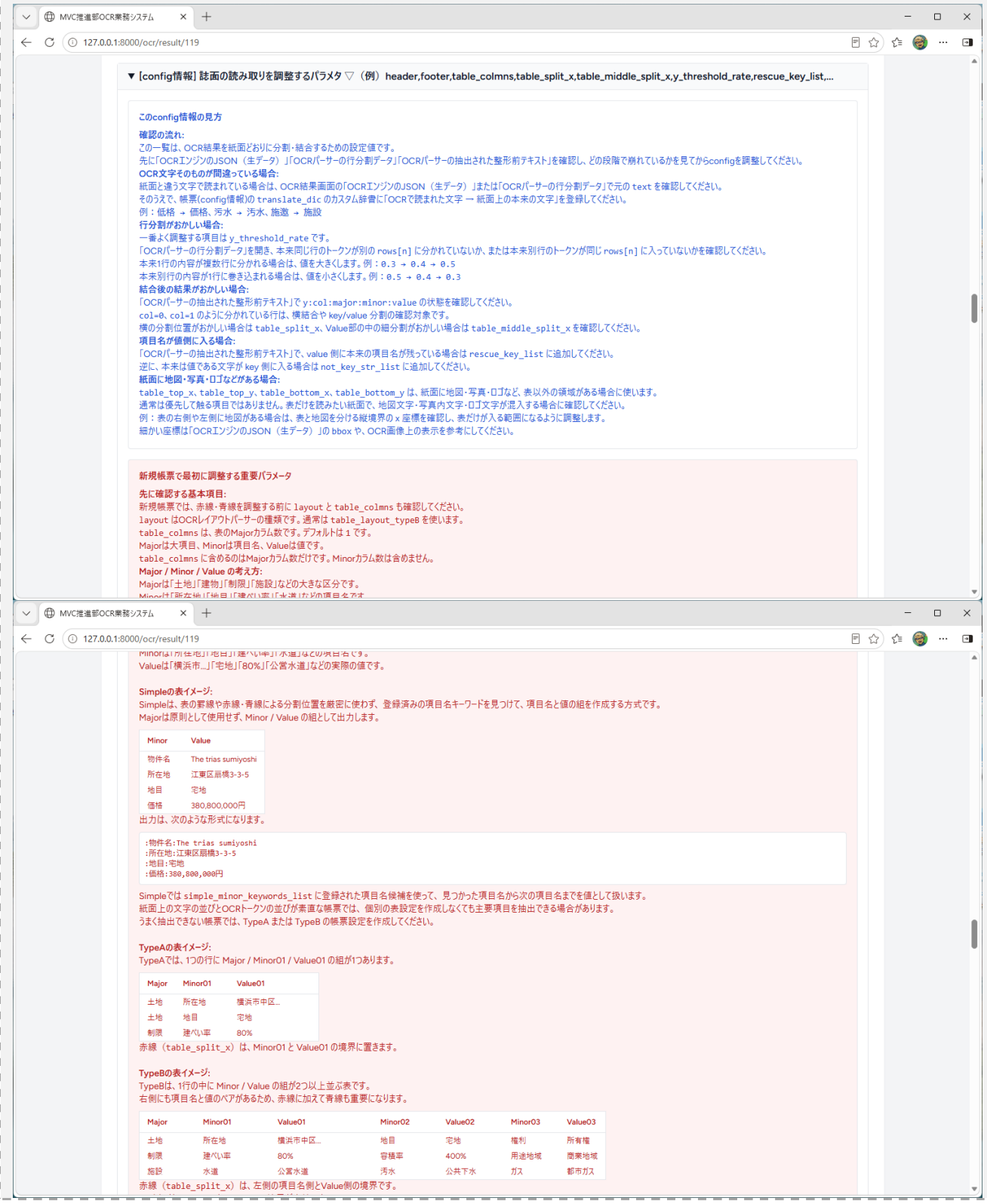
整形前テキストに誤認識された文字があり、最終結果でもそのまま残っている場合は、帳票 config の translate_dic に辞書登録を行います。整形前テキストでは揺れているが、最終結果では正しく直っている場合は、辞書変換や補正処理が正しく働いています。

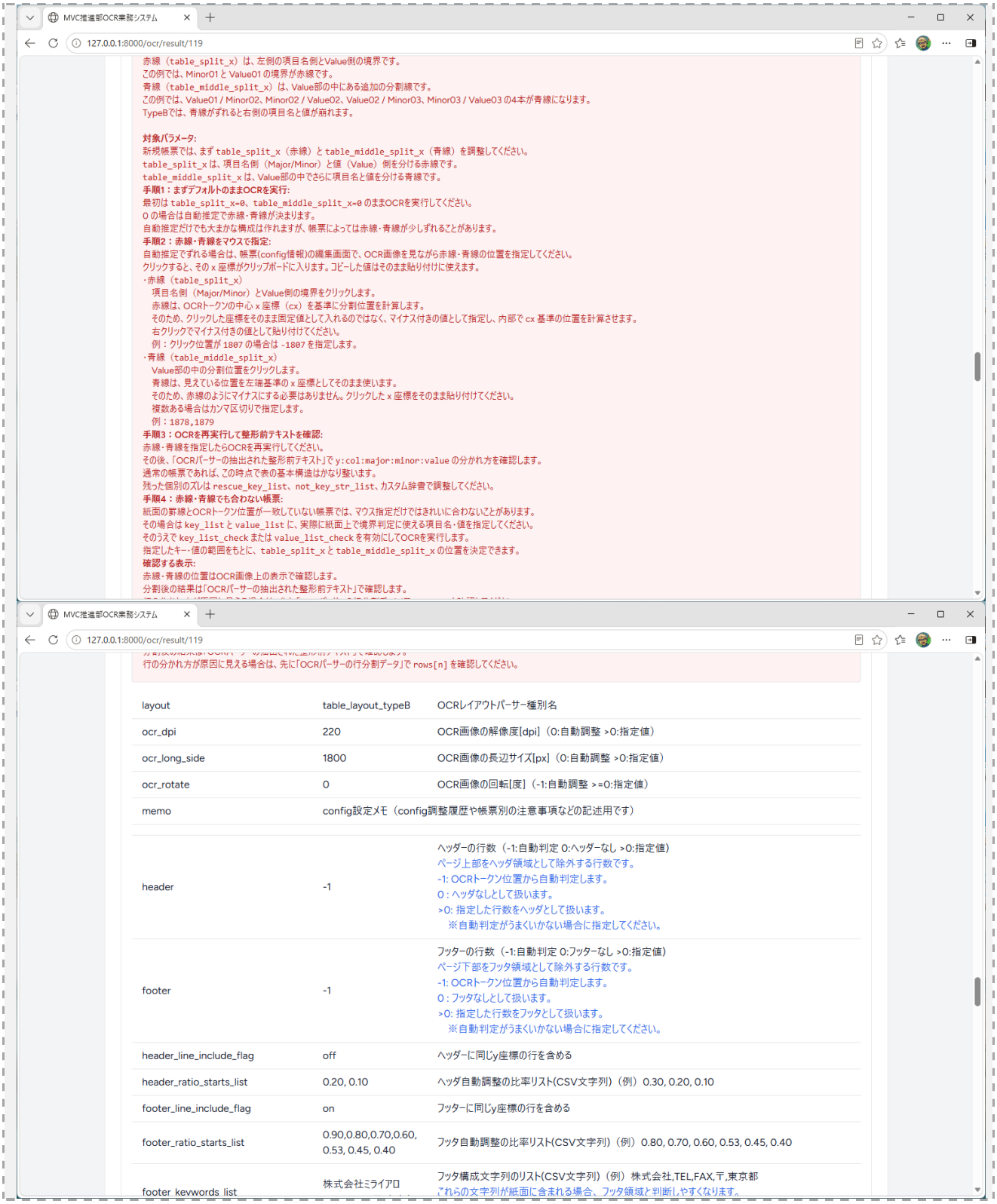
5.5 OCR 結果画面から次に進む判断

状態	操作
OCR 結果が正しい	「キーワード解析開始」ボタンをクリックします。
OCR 結果に問題がある	「帳票(config 情報)の編集画面に進む」ボタンをクリックして設定を修正します。
config 修正後に再実行する	「OCR 再読み込み画面に戻る」ボタンで戻り、OCR を再実行します。

6. 帳票 config 編集と OCR 再実行

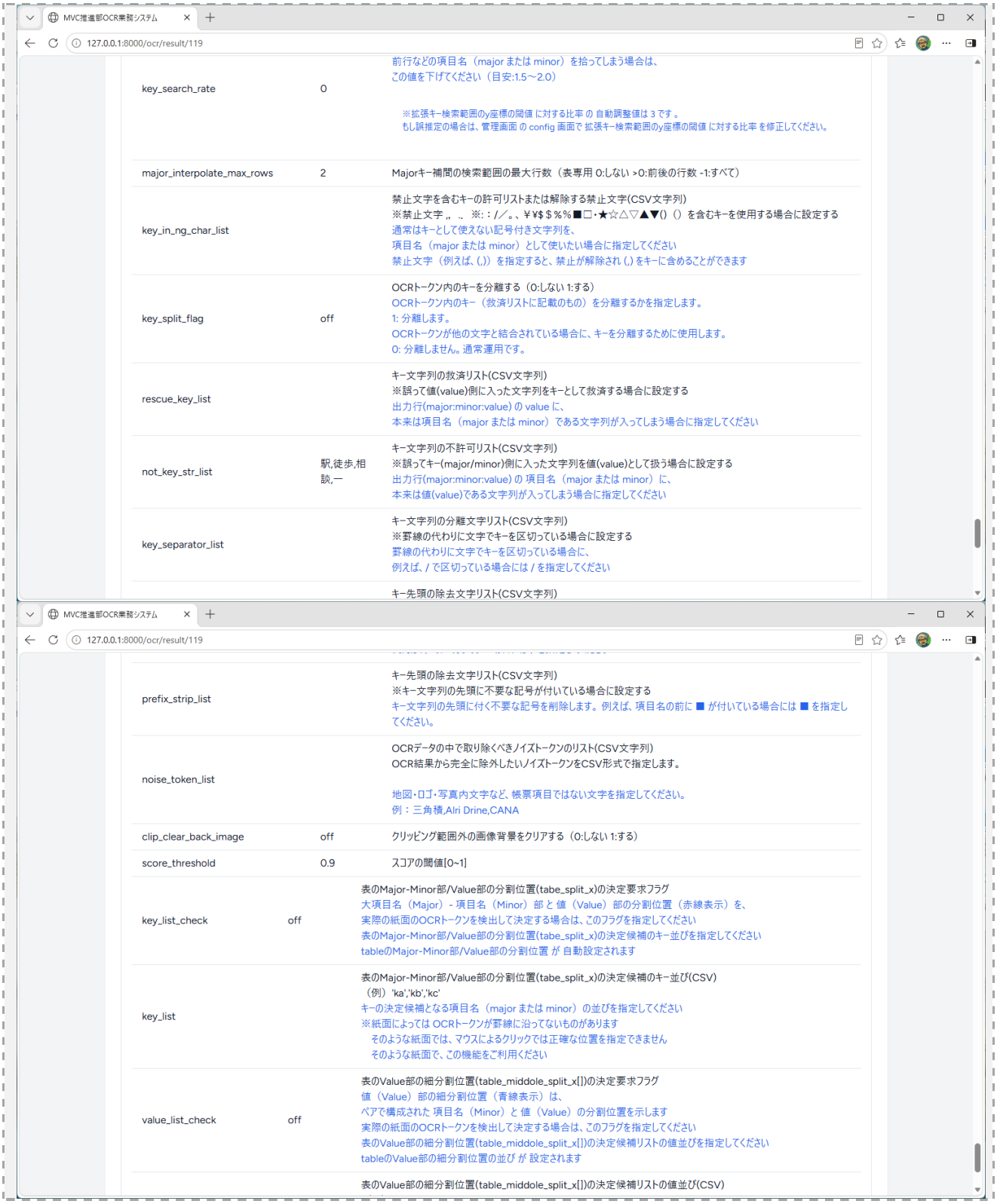
帳票 config 編集画面では、帳票ごとの読み取り設定を確認・修正できます。OCR 結果に問題がある場合は、この画面で設定を調整し、再度 OCR を実行します。


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
▼ [config情報] 誌面の読み取りを調整するパラメタ ▽（例）header,footer,table_colmns,table_split_x,table_middle_split_x,y_threshold_rate,rescue_key_list,...
このconfig情報の見方
確認の流れ:
この一覧は、OCR結果を紙面どおりに分割・結合するための設定値です。
先に「OCRエンジンのJSON（生データ）」「OCRパーサーの行分割データ」「OCRパーサーの抽出された整形前テキスト」を確認し、どの段階で崩れているかを見てからconfigを調整してください。
OCR文字そのものが間違っている場合:
紙面と違う文字で読まれている場合は、OCR結果画面の「OCRエンジンのJSON（生データ）」または「OCRパーサーの行分割データ」で元の text を確認してください。
そのうえで、帳票(config情報)の translate_dic のカスタム辞書に「OCRで読まれた文字 → 紙面上の本来の文字」を登録してください。
例：低格 → 価格、汚水 → 汚水、施邀 → 施設
行分割がおかしい場合:
一番よく調整する項目は y_threshold_rate です。
「OCRパーサーの行分割データ」を開き、本来同じ行のトークンが別の rows[n] に分かれていないか、または本来別行のトークンが同じ rows[n] に入っていないかを確認してください。
本来1行の内容が複数行に分かれる場合は、値を大きくします。例：0.3 → 0.4 → 0.5
本来別行の内容が1行に巻き込まれる場合は、値を小さくします。例：0.5 → 0.4 → 0.3
結合後の結果がおかしい場合:
「OCRパーサーの抽出された整形前テキスト」で y:col:major:minor:value の状態を確認してください。
col=0、col=1 のように分かれている行は、横結合や key/value 分割の確認対象です。
横の分割位置がおかしい場合は table_split_x、Value部の中の細分割がおかしい場合は table_middle_split_x を確認してください。
項目名が値側に入る場合:
「OCRパーサーの抽出された整形前テキスト」で、value 側に本来の項目名が残っている場合は rescue_key_list に追加してください。
逆に、本来は値である文字が key 側に入る場合は not_key_str_list に追加してください。
紙面に地図・写真・ロゴなどがある場合:
table_top_x、table_top_y、table_bottom_x、table_bottom_y は、紙面に地図・写真・ロゴなど、表以外の領域がある場合に使います。
通常は優先して触る項目ではありません。表だけを読みたい紙面で、地図文字・写真内文字・ロゴ文字が混入する場合に確認してください。
例：表の右側や左側に地図がある場合は、表と地図を分ける縦境界の x 座標を確認し、表だけが入る範囲になるように調整します。
細かい座標は「OCRエンジンのJSON（生データ）」の bbox や、OCR画像上の表示を参考にしてください。
新規帳票で最初に調整する重要パラメータ
先に確認する基本項目:
新規帳票では、赤線・青線を調整する前に layout と table_colmns も確認してください。
layout はOCRレイアウトパーサーの種類です。通常は table_layout_typeB を使います。
table_colmns は、表のMajorカラム数です。デフォルトは 1 です。
Majorは大項目、Minorは項目名、Valueは値です。
table_colmns に含めるのはMajorカラム数だけです。Minorカラム数は含めません。
Major / Minor / Value の考え方:
Majorは「土地」「建物」「制限」「施設」などの大きな区分です。
Valueは「横浜市...」「宅地」「80%」「公営水道」などの実際の値です。
Simpleの表イメージ:
Simpleは、表の罫線や赤線・青線による分割位置を厳密に使わず、登録済みの項目名キーワードを見つけて、項目名と値の組を作成する方式です。
Majorは原則として使用せず、Minor / Value の組として出力します。
Minor Value
物件名 The trias sumiyoshi
所在地 江東区扇橋3-3-5
地目 宅地
価格 380,800,000円
出力は、次のような形式になります。
:物件名:The trias sumiyoshi
:所在地:江東区扇橋3-3-5
:地目:宅地
:価格:380,800,000円
Simpleでは simple_minor_keywords_list に登録された項目名候補を使って、見つかった項目名から次の項目名までを値として扱います。
紙面上の文字の並びとOCRトークンの並びが素直な帳票では、個別の表設定を作成しなくても主要項目を抽出できる場合があります。
うまく抽出できない帳票では、TypeA または TypeB の帳票設定を作成してください。
TypeAの表イメージ:
TypeAでは、1つの行に Major / Minor01 / Value01 の組が1つあります。
Major Minor01 Value01
土地 所在地 横浜市中区...
土地 地目 宅地
制限 建ぺい率 80%
赤線（table_split_x）は、Minor01 と Value01 の境界に置きます。
TypeBの表イメージ:
TypeBは、1行の中に Minor / Value の組が2つ以上並ぶ表です。
右側にも項目名と値のペアがあるため、赤線に加えて青線も重要になります。
Major Minor01 Value01 Minor02 Value02 Minor03 Value03
土地 所在地 横浜市中区... 地目 宅地 権利 所有権
制限 建ぺい率 80% 容積率 400% 用途地域 商業地域
施設 水道 公営水道 汚水 公共下水 ガス 都市ガス
赤線（table_split_x）は、左側の項目名側とValue側の境界です。


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
赤線（table_split_x）は、左側の項目名側とValue側の境界です。
この例では、Minor01 と Value01 の境界が赤線です。
青線（table_middle_split_x）は、Value部の中にある追加の分割線です。
この例では、Value01 / Minor02、Minor02 / Value02、Value02 / Minor03、Minor03 / Value03 の4本が青線になります。
TypeBでは、青線がずれると右側の項目名と値が崩れます。
対象パラメータ:
新規帳票では、まず table_split_x（赤線）と table_middle_split_x（青線）を調整してください。
table_split_x は、項目名側（Major/Minor）と値（Value）側を分ける赤線です。
table_middle_split_x は、Value部の中でさらに項目名と値を分ける青線です。
手順1：まずデフォルトのままOCRを実行:
最初は table_split_x=0、table_middle_split_x=0 のままOCRを実行してください。
0 の場合は自動推定で赤線・青線が決まります。
自動推定だけでも大まかな構成は作れますが、帳票によっては赤線・青線が少しずれることがあります。
手順2：赤線・青線をマウスで指定:
自動推定でずれる場合は、帳票(config情報)の編集画面で、OCR画像を見ながら赤線・青線の位置を指定してください。
クリックすると、その x 座標がクリップボードに入ります。コピーした値はそのまま貼り付けに使えます。
・赤線（table_split_x）
項目名側（Major/Minor）とValue側の境界をクリックします。
赤線は、OCRトークンの中心 x 座標（cx）を基準に分割位置を計算します。
そのため、クリックした座標をそのまま固定値として入れるのではなく、マイナス付きの値として指定し、内部で cx 基準の位置を計算させます。
右クリックでマイナス付きの値として貼り付けてください。
例：クリック位置が 1807 の場合は -1807 を指定します。
・青線（table_middle_split_x）
Value部の中の分割位置をクリックします。
青線は、見えている位置を左端基準の x 座標としてそのまま使います。
そのため、赤線のようにマイナスにする必要はありません。クリックした x 座標をそのまま貼り付けてください。
複数ある場合はカンマ区切りで指定します。
例：1878,1879
手順3：OCRを再実行して整形前テキストを確認:
赤線・青線を指定したらOCRを再実行してください。
その後、「OCRパーサーの抽出された整形前テキスト」で y:col:major:minor:value の分かれ方を確認します。
通常の帳票であれば、この時点で表の基本構造はかなり整います。
残った個別のズレは rescue_key_list、not_key_str_list、カスタム辞書で調整してください。
手順4：赤線・青線でも合わない帳票:
紙面の罫線とOCRトークン位置が一致していない帳票では、マウス指定だけではきれいに合わないことがあります。
その場合は key_list と value_list に、実際に紙面上で境界判定に使える項目名・値を指定してください。
そのうえで key_list_check または value_list_check を有効にしてOCRを実行します。
指定したキー・値の範囲をもとに、table_split_x と table_middle_split_x の位置を決定できます。
確認する表示:
赤線・青線の位置はOCR画像上の表示で確認します。
分割後の結果は「OCRパーサーの抽出された整形前テキスト」で確認します。
MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
行の分かれ方が原因に見える場合は、先に「OCRパーサーの行分割データ」で rows[n] を確認してください。
layout table_layout_typeB OCRレイアウトパーサー種別名
ocr_dpi 220 OCR画像の解像度[dpi]（0:自動調整 >0:指定値）
ocr_long_side 1800 OCR画像の長辺サイズ[px]（0:自動調整 >0:指定値）
ocr_rotate 0 OCR画像の回転[度]（-1:自動調整 >=0:指定値）
memo config設定メモ（config調整履歴や帳票別の注意事項などの記述用です）
header -1 ヘッダーの行数（-1:自動判定 0:ヘッダーなし >0:指定値）
ページ上部をヘッダ領域として除外する行数です。
-1: OCRトークン位置から自動判定します。
0 : ヘッダなしとして扱います。
>0: 指定した行数をヘッダとして扱います。
※自動判定がうまくいかない場合に指定してください。
footer -1 フッターの行数（-1:自動判定 0:フッターなし >0:指定値）
ページ下部をフッタ領域として除外する行数です。
-1: OCRトークン位置から自動判定します。
0 : フッタなしとして扱います。
>0: 指定した行数をフッタとして扱います。
※自動判定がうまくいかない場合に指定してください。
header_line_include_flag off ヘッダーに同じy座標の行を含める
header_ratio_starts_list 0.20, 0.10 ヘッダ自動調整の比率リスト(CSV文字列)（例）0.30, 0.20, 0.10
footer_line_include_flag on フッターに同じy座標の行を含める
footer_ratio_starts_list 0.90,0.80,0.70,0.60, 0.53, 0.45, 0.40 フッタ自動調整の比率リスト(CSV文字列)（例）0.80, 0.70, 0.60, 0.53, 0.45, 0.40
footer_keywords_list 株式会社ミライア フッタ構成文字列のリスト(CSV文字列)（例）株式会社,TEL,FAX,〒,東京都
これらの文字列が紙面に含まれる場合、フッタ領域と判断しやすくなります。


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
footer_keywords_list
株式会社ミライアロ
ー,TEL,FAX,〒,東京都
フッタ構成文字列のリスト(CSV文字列)（例）株式会社,TEL,FAX,〒,東京都
これらの文字列が紙面に含まれる場合、フッタ領域と判断しやすくなります。
会社名は完全名称で指定してください。
remarks_list
備考
備考文字列の候補リスト(CSV文字列)（例）備考
OCR揺れで「備考」が「考」などになる紙面では、その揺れも候補に追加してください。
display_none_value_flag
off
値なし（末尾：）の通常表を表示する（0:しない 1:する）
値が空の通常行を結果に表示するかを指定します。
1: 表示します。通常運用です。紙面に存在する項目行を結果にも残ます。
0: 表示しません。Simpleテキストなどで、不要な値なし行を削除する場合に指定します。
remark_rollback_empty_value_lines
on
直前の値なし（末尾：）の通常行も、備考として吸収する（表専用 0:しない 1:する）
備考開始直前にある「値なし通常行」も備考へ吸収するかを指定します。
0: 吸収しません。
1: 吸収します。
備考欄は 複数行にまたがって、配置されることが多く、
備考キーが中央付近に印字される紙面では、前方数行が bullet(::value) 行や
value なし通常行として出力される場合があります。
この設定を有効にすると、それらの行を備考本文へ吸収します。
ただし、本文の通常項目まで誤って備考へ吸収する場合があるため、
備考吸収の最大行数と合わせて調整してください。
remark_max_rollback_rows
5
直前の行を備考として吸収(ロールバック)するかの最大行数（表専用 0:しない >0:指定行数 -1:すべて）
備考開始直前の行を、最大何行まで備考へ吸収するかを指定します。
0: 吸収しません。
>0: 指定した行数まで吸収します。
-1: 直前の吸収可能行をすべて吸収します。
備考欄は 複数行にまたがって、配置されることが多ため、備考前方の bullet(::value) 行や value なし通常行を備考へ吸収しないと、通常は紙面の備考内容と一致しません。
備考の吸収 は、値のある通常行（minor:value）に到達すると停止します。
simple_minor_keywords_list
Simpleテキストの項目名(minor)候補のリスト(CSV文字列) (正規表現も指定可)
（例）価格,住所,交通
table_top_x
-1
tableの左上隅のx座標[px]（表専用 -1:自動設定(0) >=0:指定値）
table_top_y
-1
tableの左上隅のy座標[px]（表専用 -1:自動設定(0) >=0:指定値）
table_bottom_x
-1
tableの右下隅のx座標[px]（表専用 -1:自動設定(紙面幅) <0:紙面幅-指定値 >=0:指定値）
table_bottom_y
-1
tableの右下隅のy座標[px]（表専用 -1:自動設定(紙面高) <0:紙面高-指定値 >=0:指定値）
translate_dic
中国語フォント変換辞書のカスタム定義(JSON)（例）{"word_variant_map": {"洪谷区":"渋谷区","所有推":"所有権",...}}
{
"word_variant_map": []
}
・カスタム辞書にはこの帳票で「OCRにより間違って読み取られた言葉 → 正しい言葉」の形で登録してください
・正しい言葉をさらに別の言葉に変えるような登録はしないでください
・登録する言葉は、できるだけ意味がはっきりした単語で登録してください
・短すぎる言葉（2〜3文字）は、思わぬ影響が出るため基本的に登録しないでください
・置き換えた後に、元の言葉が残らないようにしてください （同じ変換が何度も繰り返されてしまう原因になります）
・うまく直らない場合は、長い文章のまま登録しても問題ありません
・新しく追加した場合は、必ず実際の帳票で結果を確認してください
table_colmns
1
tableのMajorカラム数 (表専用 0:自動設定 >0:指定値)（デフォルト値:1）
大項目名（Major）のカラム数を指定して下さい
※項目名（Minor）のカラム数は含みません
table_split_x
366
tableのMajor-Minor部/Value部の分割位置[px]（表専用 <0:自動推定 0:自動調整 >0:指定値）
大項目名（Major）- 項目名（Minor）部 と 値（Value）部の分割位置（赤線表示）は、
分割位置の罫線の上でマウスを左クリックして x座標を拾ってください（クリップボードに入ります）
拾った x座標を、右クリックで -（マイナス）を付けて指定してください（ペーストされます）
tableのMajor-Minor部/Value部の分割位置 が 自動設定されます
table_middle_split_x
713,942
tableのValue部の細分割位置の並び[px](CSV文字列)（表専用 0:自動設定 >0:指定値1,指定値2,,,指定値[Major-Minor部以降のカラム数-1]）
値（Value）部の細分割位置（青線表示）は、
ペアで構成された 項目名（Minor）と 値（Value）の分割位置を示します


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
table_middle_split_x
713,942
ペアで構成された 項目名（Minor）と 値（Value）の分割位置を示します
分割位置の罫線の上でマウスを左クリックして x座標を拾ってください（クリップボードに入ります）
拾った x座標を、,（カンマ）区切りで右クリックして指定してください（ペーストされます）
tableのValue部の細分割位置の並び が 設定されます
tabel_major_minor_reverse
off
table(Matrix)のキー順序(Major-Minor)
（表専用 0: 行-列(Major-Minor)フラグ 1: 行-列(Minor-Major)
y_threshold_row_anchor_mode
0
行分割アルゴリズム（-1:自動調整 0:直前 token 基準 1:行先頭 token 基準 2:行内中央値(可変レンジ) 基準 3:局所行間ピッチ追従 基準）
x_threshold
0
x座標の閾値[px]（0:自動調整 >0:指定値）（デフォルト値:10）
y_threshold
0
y座標の閾値[px]（-1:自動推定 0:自動調整 >0:指定値）（デフォルト値:27）
※y座標の閾値 の 自動調整値は 14 です 。
もし誤推定の場合は、管理画面 の config 画面で y座標の閾値 を修正してください。
y_threshold_rate
0.4
y座標の中央値の閾値に対する比率（0:自動調整 >0:指定値）（デフォルト値:0.6)
・確認場所
「OCRパーサーの行分割データ」を開き、同じ rows[n] に入るべきか、別の rows[n] に分かれるべきかを確認してください。
・本来1行の内容が複数の rows[n] に分かれている場合
この値を大きくしてください。
例：0.3 → 0.4 → 0.5
・本来別行の内容が同じ rows[n] に巻き込まれている場合
この値を小さくしてください。
例：0.5 → 0.4 → 0.3
・同じトークンや同じ項目が複数行に重複して出る場合
行分割の範囲が広すぎる可能性があります。まずこの値を小さくして確認してください。
・調整しても改善しない場合
行分割アルゴリズム（y_threshold_row_anchor_mode）を変更してから、再度この値を調整してください。
・自動推定したい場合
y_threshold=-1 を指定して、y座標の閾値を自動推定してください。
y_threshold_same_row_hori_rate
0
同一行判定（横結合）の y座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.0）
※同一行判定（横結合）のy座標の閾値 に対する比率 の 自動調整値は 1.1 です 。
もし誤推定の場合は、管理画面 の config 画面で 同一行判定（横結合）のy座標の閾値 に対する比率 を修正してください。
y_threshold_same_row_vert_rate
1.4
同一行判定（縦結合）の y座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.7）
y_threshold_blocktext_rate
0
block_text判定の y座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.2）
key_max_length
0
tableのキーの有効最大文字数 (表専用 -1:制限なし 0:自動設定 >0:指定値)（デフォルト値:6）
key_x_threshold
0
tableのトークン結合判定 (表専用 0:自動調整 -1:文字数指定 >0:px指定値）（デフォルト値:-1）
key_concat_threshold
0
tableのトークン結合最大文字数（表専用 0:自動調整 指定値(整数:n, 実数:m.n, 実数:k.k)）
※ n:最大文字数 m:最小文字数 k:対象文字数（デフォルト値:1）
key_minor_threshold
0
tableの項目名数(block_text構造の閾値 (表専用 0:自動設定 >0:指定値)（デフォルト値:1）
block_text_value_x_limit_threshold
0
block_text構造のバリュー探索の最大距離[px]（表専用 0:自動設定 >0:指定値）（デフォルト値:650）
block_minor_outlier_dx_threshold
0
block_text構造の左方向外れ項目名の判定距離[px]（表専用 0:自動設定 >0:指定値）（デフォルト値:120）
key_search_mode
3
キー検索アルゴリズム（0:自動調整 1:行内探索 2:拡張探索（valueのみ行でも外側キーを探す） 3:（value-only行では外側キーを探さない））
拡張キー検索範囲のy座標の閾値に対する比率（0:自動調整 >0:指定値）（デフォルト値:3.0）
直前の備考に入るべき行が bullet(::Value)行にならず、
前行などの項目名（major または minor）を拾ってしまう場合は、
この値を下げてください（目安:1.5～2.0）


MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
key_search_rate
0
前行などの項目名（major または minor）を拾ってしまう場合は、
この値を下げてください（目安:1.5～2.0）
※拡張キー検索範囲のy座標の閾値 に対する比率 の 自動調整値は 3 です 。
もし誤推定の場合は、管理画面 の config 画面で 拡張キー検索範囲のy座標の閾値 に対する比率 を修正してください。
major_interpolate_max_rows
2
Majorキー補間の検索範囲の最大行数（表専用 0:しない >0:前後の行数 -1:すべて）
key_in_ng_char_list
禁止文字を含むキーの許可リストまたは解除する禁止文字(CSV文字列)
※禁止文字 ,, ．． ※: : / ／ 。、￥¥$＄%％■□・★☆△▽▲▼() （）を含むキーを使用する場合に設定する
通常はキーとして使えない記号付き文字列を、
項目名（major または minor）として使いたい場合に指定してください
禁止文字（例えば、(,)）を指定すると、禁止が解除され (,) をキーに含めることができます
key_split_flag
off
OCRトークン内のキーを分離する（0:しない 1:する）
OCRトークン内のキー（救済リストに記載のもの）を分離するかを指定します。
1: 分離します。
OCRトークンが他の文字と結合されている場合に、キーを分離するために使用します。
0: 分離しません。通常運用です。
rescue_key_list
キー文字列の救済リスト(CSV文字列)
※誤って値(value)側に入った文字列をキーとして救済する場合に設定する
出力行(major:minor:value) の value に、
本来は項目名（major または minor）である文字列が入ってしまう場合に指定してください
not_key_str_list
駅,徒歩,相談,一
キー文字列の不許可リスト(CSV文字列)
※誤ってキー(major/minor)側に入った文字列を値(value)として扱う場合に設定する
出力行(major:minor:value) の 項目名（major または minor）に、
本来は値(value)である文字列が入ってしまう場合に指定してください
key_separator_list
キー文字列の分離文字リスト(CSV文字列)
※罫線の代わりに文字でキーを区切っている場合に設定する
罫線の代わりに文字でキーを区切っている場合に、
例えば、/ で区切っている場合には / を指定してください
キー先頭の除去文字リスト(CSV文字列)
MVC推進部OCR業務システム
127.0.0.1:8000/ocr/result/119
prefix_strip_list
キー先頭の除去文字リスト(CSV文字列)
※キー文字列の先頭に不要な記号が付いている場合に設定する
キー文字列の先頭に付く不要な記号を削除します。例えば、項目名の前に ■ が付いている場合には ■ を指定してください。
noise_token_list
OCRデータの中で取り除くべきノイズトークンのリスト(CSV文字列)
OCR結果から完全に除外したいノイズトークンをCSV形式で指定します。
地図・ロゴ・写真内文字など、帳票項目ではない文字を指定してください。
例：三角積,Alri Drine,CANA
clip_clear_back_image
off
クリッピング範囲外の画像背景をクリアする（0:しない 1:する）
score_threshold
0.9
スコアの閾値[0~1]
key_list_check
off
表のMajor-Minor部/Value部の分割位置(tabe_split_x)の決定要求フラグ
大項目名（Major）- 項目名（Minor）部 と 値（Value）部の分割位置（赤線表示）を、
実際の紙面のOCRトークンを検出して決定する場合は、このフラグを指定してください
表のMajor-Minor部/Value部の分割位置(tabe_split_x)の決定候補のキー並びを指定してください
tableのMajor-Minor部/Value部の分割位置 が 自動設定されます
key_list
表のMajor-Minor部/Value部の分割位置(tabe_split_x)の決定候補のキー並び(CSV)
（例）'ka','kb','kc'
キーの決定候補となる項目名（major または minor）の並びを指定してください
※紙面によっては OCRトークンが罫線に沿ってないものがあります
そのような紙面では、マウスによるクリックでは正確な位置を指定できません
そのような紙面で、この機能をご利用ください
value_list_check
off
表のValue部の細分割位置(table_middole_split_x[])の決定要求フラグ
値（Value）部の細分割位置（青線表示）は、
ペアで構成された 項目名（Minor）と 値（Value）の分割位置を示します
実際の紙面のOCRトークンを検出して決定する場合は、このフラグを指定してください
表のValue部の細分割位置(table_middole_split_x[])の決定候補リストの値並びを指定してください
tableのValue部の細分割位置の並び が 設定されます
表のValue部の細分割位置(table_middole_split_x[])の決定候補リストの値並び(CSV)



図：帳票 config 編集画面

6.1 config 編集後の再 OCR の流れ

順序	操作
1	OCR 結果画面で問題箇所を確認します。
2	「帳票(config 情報)の編集画面に進む」をクリックします。
3	帳票 config 編集画面で必要な設定を修正し、保存します。
4	左メニューの「MVC 推進部 OCR 業務システムに戻る」をクリックします。
5	OCR 結果画面で「OCR 再読み込み画面に戻る」をクリックします。
6	OCR 再読み込み画面で再度 OCR を実行します。
7	再実行後の OCR 結果を確認し、問題なければ「キーワード解析開始」をクリックします。

6.2 config 編集画面の主な項目

画面操作系の主な項目

項目	説明
画像プレビュー	読み取り対象の紙面画像を表示します。表組（TypeA / TypeB）では赤線・青線が表示されます。Simple では基本的に赤線・青線は使用しません。
クリック座標取得	画像上をクリックすると座標を取得できます。X 座標はクリップボードへ入り、設定欄へ貼り付けできます。赤線・青線や表範囲を指定する場合に使用します。
帳票 config 設定欄	OCR レイアウト種別、ヘッダー／フッター、表の分割位置、キーワード、辞書など、帳票ごとの読み取り設定を編集します。
保存 / キャンセル	設定を変更した場合は保存します。保存後、OCR 結果画面へ戻り、再 OCR を実行して結果を確認します。
MVC 推進部 OCR 業務システムに戻る	管理画面から通常の OCR 業務システム画面へ戻ります。config 編集後の確認・再 OCR のた

	めに使用します。

主な config パラメータ

パラメータ名	説明
ocr_dpi / ocr_long_side / ocr_rotate	OCR 画像処理系の設定です。Simple ひな型を流用する場合は、帳票ごとに自動設定へ戻すことが重要です。
memo	調整履歴、帳票別の注意事項、標準設定から変更した理由を残します。
header	ヘッダーの行数を指定します（-1: 自動判定、0: ヘッダーなし、>0: 指定行数）。ページ上部をヘッダー領域として除外します。自動判定がうまくいかない場合に指定します。
footer	フッターの行数を指定します（-1: 自動判定、0: フッターなし、>0: 指定行数）。ページ下部をフッター領域として除外します。自動判定がうまくいかない場合に指定します。
remarks_list / remark_rollback_empty_value_lines / remark_max_rollback_rows	備考欄の開始判定、ロールバック行数、値なし行の吸収などを調整します。備考本文が途中で通常行に戻る場合や、備考前方の行を備考に含めたい場合に確認します。
simple_minor_keywords_list	Simple で使用する項目名（minor）候補のリストです。キーを追加する場合は語句を追加し、本来値にしたい語句がキーになる場合は -所有権 のようにマイナス指定で解除します。
table_top_x / table_top_y / table_bottom_x / table_bottom_y	表の解析対象範囲を指定します。右図版、左図版、地図、写真、ロゴなどがある紙面で、表だけを対象にしたい場合に使います。
table_split_x	表組（TypeA / TypeB）の Major-Minor 部 / Value 部 の分割位置です。画像上では赤線で表示されます。分割位置の罫線上を左クリックして X 座標を取得し、右クリックで -（マイナス）を付けて指定すると、自動推定の基準にできます。
table_middle_split_x	Value 部の細分割位置の並びです。画像上では青線で表示されます。1 行に複数の「項目名（Minor）と値（Value）」のペアがある場合に、Value 部を区切るために指定します。複数指定する場合は 664,782 のようにカンマ区切りで入力します。
key_list / value_list	通常調整で合わない場合に、キーや値を指定して赤線・青線を推定させるための補助設定

	です。マウスによる赤線・青線の設定ができない紙面で使用します。
key_in_ng_char_list	(,), (,) ,/,／など、通常はキーとして使えない禁止文字を、項目名（major または minor）として使いたい場合に許可する設定です。禁止文字には ,, .. ※: : /／。、￥¥$ $ %％■□・★☆△▽▲▼() （）などがあります。たとえば (,) を指定すると、括弧を含むキーを使用できます。
key_split_flag	OCR トークン内のキーを分離するかを指定します（off/0: 分離しない、on/1: 分離する）。OCR トークンが他の文字と結合されている場合に、rescue_key_list に記載したキーを分離するために使用します。通常運用では off です。
rescue_key_list	キー文字列の救済リストです。誤って値（value）側に入った文字列を、本来の項目名（major または minor）として救済する場合に指定します。出力行 major:minor:value の value に、本来キーである文字列が混入する場合に使用します。
not_key_str_list	キー文字列の不許可リストです。誤ってキー（major/minor）側に入った文字列を値（value）として扱う場合に指定します。本来値である語句がキーになって値が空になる場合に使用します。
key_separator_list	:, : など、キー文字列を分離する文字のリストです。罫線の代わりに文字でキーを区切っている場合に指定します。たとえば / で区切っている場合には / を指定します。Simple と表組のどちらでも使用します。
noise_token_list	OCR データの中で完全に除外したいノイズトークンのリストです。地図、ロゴ、写真内文字など、帳票項目ではない文字を CSV 形式で指定します。例: 三角積,Alri Drine,CANA。Simple と表組のどちらでも使用します。
translate_dic	OCR 文字揺れ・誤認識を補正する辞書です。OCR により間違って読み取られた語句を、帳票 config 内の word_variant_map で正しい語句へ変換します。

注意：キーワード系設定の登録ルール

帳票 config で、キーワード名・項目名を完全一致または候補として指定する項目は、Simple に限らず、原則として文字の途中に空白を入れずに登録します。対象例は、Simple キーワード、備考文字列（remarks_list）、rescue_key_list、not_key_str_list、key_separator_list などです。紙面上で「備 考」「所在 地」「価 格」のように文字間に半角スペース・全角スペースが見える場合でも、登録値は「備考」「所在地」「価格」のように空白なしにしてください。表組（TypeA）/ 表組（TypeB） などの表組みパーサーでは従来からキー候補を結合して判定しており、Simple も同じ考え方で比較します。

not_key_str_list などでは、括弧付き項目名の括弧内だけを指定する場合があります。その場合も、指定する項目名部分の途中には空白を入れません。空白を残すと、OCR トークンの分割や文字間スペースの入り方によって、同じ項目名として判定できない場合があります。

6.2.1 Simple（テキスト）レイアウトの調整方法

Simple（テキスト）レイアウトは、赤線・青線を使わず、OCR 結果の文字列と Simple 用キーワードリストをもとに、Minor / Value の組を作成します。抽出結果が崩れる場合は、最終結果だけで判断せず、OCR 結果画面の折りたたみ情報から「OCR パーサーの抽出された整形前テキスト」を開き、どの段階で崩れているかを確認します。

確認に使う主な情報

確認する情報	見る内容	主な調整
OCR パーサーの抽出された整形前テキスト	y:col:major:minor:value の形式で、最終整形前の行を確認します。	Simple テキストの項目名(minor)候補のリスト、マイナスキーワード、translate_dic
OCR パーサーの行分割データ	rows[n] 単位で、本来別行のトークンが同じ行に混ざっていないかを確認します。	y_threshold_rate、y_threshold_row_anchor_mode
抽出されたテキスト	最終的な major:minor:value の出力結果を確認します。	再 OCR 後の確認、キーワード解析結果との照合

本来キーになるべき語句がキーになっていない場合

整形前テキストで、次のような value-only 行（bullet 行）を確認します。

```
629.0:0:::都市計画区域 市街化区域
631.0:0:::接道 南側 私道 幅員約 4.0m
```

この形式は、キーが取れず、値だけの行として出力されている状態です。値の先頭トークンが本来は項目名キーである場合は、帳票 config の「Simple テキストの項目名(minor)候補のリスト」に、そのキーを追加します。

```
simple_minor_keywords_list 追加例:
都市計画区域
接道
```

追加後に OCR を再実行し、次のようにキーと値に分かれることを確認します。

```
:都市計画区域:市街化区域
:接道:南側 私道 幅員約 4.0m
```

本来値になるべき語句がキーになっている場合

整形前テキストで、次のようなキーだけの行を確認します。

```
517.0:1::権利:
517.0:2::所有権:
```

このように、本来は値として扱うべき語句がキーになり、値が空になっている場合は、帳票 config の「Simple テキストの項目名(minor)候補のリスト」にマイナスキーワードとして指定します。

```
simple_minor_keywords_list 追加例:
-所有権
```

マイナスキーワードを指定すると、その帳票では該当語句を Simple キーとして扱わなくなります。これにより、本来 Value になるべき語句がキー化して、値が出力されない問題を避けられます。

```
修正前:
:所有権:

修正後:
:権利:所有権
```

OCR 文字揺れが原因の場合

キーにならない原因が OCR 文字揺れの場合は、揺れた文字列を Simple キーワードに追加しません。先に translate_dic で正しい語句へ補正し、正しい語句を Simple キーワードとして扱います。

```
辞書登録例:
接进 → 接道
引波 → 引渡
月策 → 月築
```

たとえば、紙面上は「接道」なのに OCR で「接进」と読まれている場合、「接进」を Simple キーワードに追加するのではなく、translate_dic で「接道」に補正します。

Simple で使うキー認識・値振り分けの補助設定

Simple（テキスト）レイアウトでも、キーの認識と値への振り分けを補助するために、次の設定を使用します。Simple は赤線・青線を使いませんが、OCR トークンの中にキーが含まれているか、キーとして使ってよい文字か、値側に入ってしまった語句をキーとして救済するか、という判定にはこれらの設定が効きます。

key_in_ng_char_list：括弧やスラッシュなど、通常はキーとして使えない文字を含む項目名をキーとして使いたい場合に指定します。例として、(住居表示) や （公簿） のような括弧付き項目をキーとして扱いたい場合に使用します。

key_split_flag：OCR トークン内に rescue_key_list のキーが他の文字と結合している場合、そのキーを分離するかを指定します。通常は off ですが、「価格 1,000 万円」のようにキーと値が 1 つの OCR トークンに結合している場合などに、on を検討します。

rescue_key_list：Simple で value-only 行（::値）になった行や、値の途中に本来キーである語句が混入した場合に、その語句をキーとして救済するために指定します。Simple キーワードに追加するだけでなく、トークン内から分離してキー化したい場合に使用します。

key_separator_list：罫線ではなく「:」「：」「/」などの文字でキーと値を区切っている帳票で使用します。Simple でも、文字列の中に区切り文字がある場合にキーと値を分ける補助になります。

noise_token_list：地図・ロゴ・写真内文字など、帳票項目ではない OCR トークンを完全に除外したい場合に指定します。Simple では、不要トークンが value-only 行として残り、交通や備考などへ誤マップされることを防ぐために使用します。

行分割が原因の場合

本来は別行のトークンが同じ rows[n] に混ざっている場合、Simple キーワードを追加しても正しく分離できないことがあります。この場合は、OCR パーサーの行分割データを確認し、y_threshold_rate を調整します。

現象	確認する内容	主な調整
本来別行の文字列が同じ行に混ざる	rows[n] の中に上下 2 段のトークンが混在している。	y_threshold_rate を小さくします。
本来 1 行の文字列が複数行に分かれる	1 つの項目・値が複数の rows[n] に分断されている。	y_threshold_rate を大きくします。
y_threshold_rate で改善しない	行の基準点が帳票に合っていない。	y_threshold_row_anchor_mode を変更して再調整します。

フッター・備考が混入する場合

会社名、担当者、TEL、FAX などが本文の value-only 行として残り、交通や備考などへ誤マップされる場合は、footer または footer_keywords_list を調整します。逆に、備考本文までフッターに吸収される場合は、footer_ratio_starts_list や footer の指定を見直します。

Simple で追いすぎない判断

Simple は、紙面の文字並びが素直な帳票を簡易的に抽出するための方式です。項目名が行方向に結合されている、値がキーの前にある、OCR トークンの並びが紙面と大きく異なる、大きな金額文字や面積文字が通常表と別扱いになっている場合は、Simple で無理に調整せず、表組（TypeA）または表組（TypeB）を使用してください。

6.2.2 表組（TypeA / TypeB）レイアウトの調整方法

表組（TypeA）/ 表組（TypeB）レイアウトでは、紙面画像上の赤線・青線を使って、表のどこまでがキー部で、どこからが値部かを指定します。Simple と異なり、表の列構造を明示して解析するため、行・列が整理された帳票や、正確な分解が必要な帳票に向いています。

線	config 項目	役割
赤線	table_split_x	Major / Minor 部と Value 部の境界を指定します。
青線	table_middle_split_x	Value 部の中に複数の Minor / Value ペアがある場合、その区切り位置を指定します。複数ある場合はカンマ区切りで指定します。
表範囲	table_top_x, table_top_y, table_bottom_x, table_bottom_y	表以外の図面・余白・フッターなどを除外するために、解析対象範囲を指定します。

赤線の調整方法

赤線は、表の左側にある大項目・項目名（Major / Minor）と、右側にある値（Value）の境界です。
OCR 結果で、値がキー側に入る、またはキーが値側に入る場合は、赤線位置を見直します。

調整対象:
`table_split_x`

画像プレビュー上で境界となる罫線付近をクリックし、取得した X 座標を table_split_x に設定します。
紙面によっては、見た目の罫線と OCR トークン位置が少しずれるため、OCR 結果を見ながら再調整します。

青線の調整方法

青線は、Value 部の中に複数の項目ペアが並ぶ帳票で使用します。1 行の中に「項目名 1 値 1 項目名 2 値 2」のような構造がある場合、青線で Value 部を区切ることで、右側の項目名と値を分けやすくします。

調整対象:
`table_middle_split_x`

複数指定例:
`664, 782`

青線が不足していると、右側の項目名が値に混入したり、複数の値が 1 つの Value に連結されたりします。青線が多すぎる、または位置がずれている場合は、本来 1 つの値が分断されることがあります。

表組で使うキー認識・救済系の調整手順

表組（TypeA / TypeB）では、まず赤線（table_split_x）と青線（table_middle_split_x）で表の列構造を合わせます。ただし、赤線・青線が合っていても、OCR トークンの結合状態や禁止文字の扱いにより、本来キーになる語句が値側に入ったり、本来値になる語句がキー側に立ったりすることがあります。
その場合は、抽出されたテキストとキーワード解析結果を確認し、必要に応じて rescue_key_list、not_key_str_list、key_split_flag、key_in_ng_char_list を併用します。

確認に使う主な情報

確認する情報	見る内容	主な調整
抽出されたテキスト	major:minor:value のどこに本来キー・本来値が出ているかを確認します。	rescue_key_list、not_key_str_list、key_split_flag、key_in_ng_char_list
キーワード解析結果	Value が対応する Excel 項目に正しく入っているか、未設定・重複・誤マップがないかを確認します。	Excel 項目選択、または再 OCR 前提の config 調整
OCR パーサーの抽出された整形前テキスト	最終結果になる前の段階で、キーと値がどこで崩れているかを確認します。	赤線・青線、rescue_key_list、not_key_str_list

本来キーになるべき語句が値側に入っている場合

抽出されたテキストで、出力行 major:minor:value の value の中に、本来は項目名（major または minor）として扱うべき語句が混入していないか確認します。

例:

土地:所在地:東京都品川区・・・ 交通 JR 各線「品川」駅 徒歩 13 分

この例では、「交通」が本来キーであるにもかかわらず、value 側に入っています。この場合は、帳票 config の rescue_key_list にキー語句を指定します。

rescue_key_list 追加例:

交通

OCR トークン内でキーと値が 1 つに結合している場合は、key_split_flag を on にすると、rescue_key_list に記載したキーをトークン内から分離しやすくなります。

本来値になるべき語句がキー側に立っている場合

抽出されたテキストで、本来は値として扱うべき語句が major または minor に出て、値が空になっていないか確認します。添付書類、会社名、帳票内の説明語句などがキーとして立つ場合があります。

例:

:公図:
:建物図面:

この例では、「公図」「建物図面」は項目名ではなく、添付書類などの値として扱うべき語句です。この場合は、帳票 config の not_key_str_list に指定して、キーとして扱わないようにします。

not_key_str_list 追加例:

公図,建物図面

not_key_str_list は、誤ってキー（major/minor）側に入った文字列を、値（value）として扱うための設定です。Simple のマイナスキーワードと同じく、「本来値なのにキーになっている」場合に使います。

禁止文字を含むキーを使いたい場合

括弧、スラッシュ、コロンなどを含む語句は、通常はキーとして使えない場合があります。紙面上の項目名が (住居表示)、（公簿）、地積（実測）、交通/アクセス のように記号を含む場合は、key_in_ng_char_list でその禁止文字を許可します。

key_in_ng_char_list 指定例:

(,),（,）,/,／

文字区切り・ノイズが原因の場合

罫線ではなく「:」「：」「/」などの文字でキーと値を区切っている帳票では、key_separator_list を指定します。地図・ロゴ・写真内文字など、帳票項目ではない OCR トークンが混入する場合は、noise_token_list で完全に除外します。

key_separator_list 例:

:, :

noise_token_list 例:

三角積,Alri Drine,CANA

TypeA と TypeB の使い分け

TypeA は、1 行の中に基本的に 1 組の「項目名（Minor）と値（Value）」がある表に向いています。
TypeB は、1 行の中に複数の「項目名（Minor）と値（Value）」のペアが横に並ぶ表に向いています。
TypeB では、赤線に加えて青線（table_middle_split_x）を使い、右側のペアを分離します。

レイアウト	向いている帳票	主な調整
TypeA	1 行につき、基本的に 1 組の項目名と値が並ぶ帳票。	赤線 table_split_x を中心に調整します。
TypeB	1 行の中に複数の項目名・値ペアが横並びになる帳票。	赤線 table_split_x と青線 table_middle_split_x を調整します。

表範囲の調整

表の右側または左側に図面がある帳票、余白やフッターに不要な文字が多い帳票では、表範囲を指定して解析対象を絞ります。表範囲を指定する場合は、画像プレビューで表の左上・右下の座標を確認し、table_top_x、table_top_y、table_bottom_x、table_bottom_y を設定します。

調整後の確認

赤線・青線・表範囲を変更したら、必ず OCR を再実行し、抽出されたテキストとキーワード解析結果を確認します。確認時は、見た目の Major / Minor ではなく、最終的に「対応する Excel 項目名」に正しい値が入っているかを基準にします。

確認項目	見る内容
抽出されたテキスト	major:minor:value が紙面の内容と対応しているか。
整形前テキスト	キーと値がどの段階で崩れているか。
キーワード解析結果	対応する Excel 項目名と Value が正しいか。
CSV / Excel 出力	テスト結果と本番結果が完全一致するか。

6.3 AI による OCR 文字揺れ辞書作成

帳票 config 編集画面を開いているときは、管理画面左メニューの「AI による OCR の揺れ対応の辞書作成」を使用できます。これは、現在表示している帳票 config の情報をもとに、OCR 文字揺れに対応する辞書作成を AI で補助する機能です。

この機能は、OCR 結果に中国語風の文字、文字化け、誤読語句が残っている場合に使用します。帳票 config のカスタム辞書（translate_dic）へ、この帳票でのみ有効な word_variant_map を登録し、OCR 再実行時に誤認識語句を正しい語句へ変換します。

操作の流れは次のとおりです。

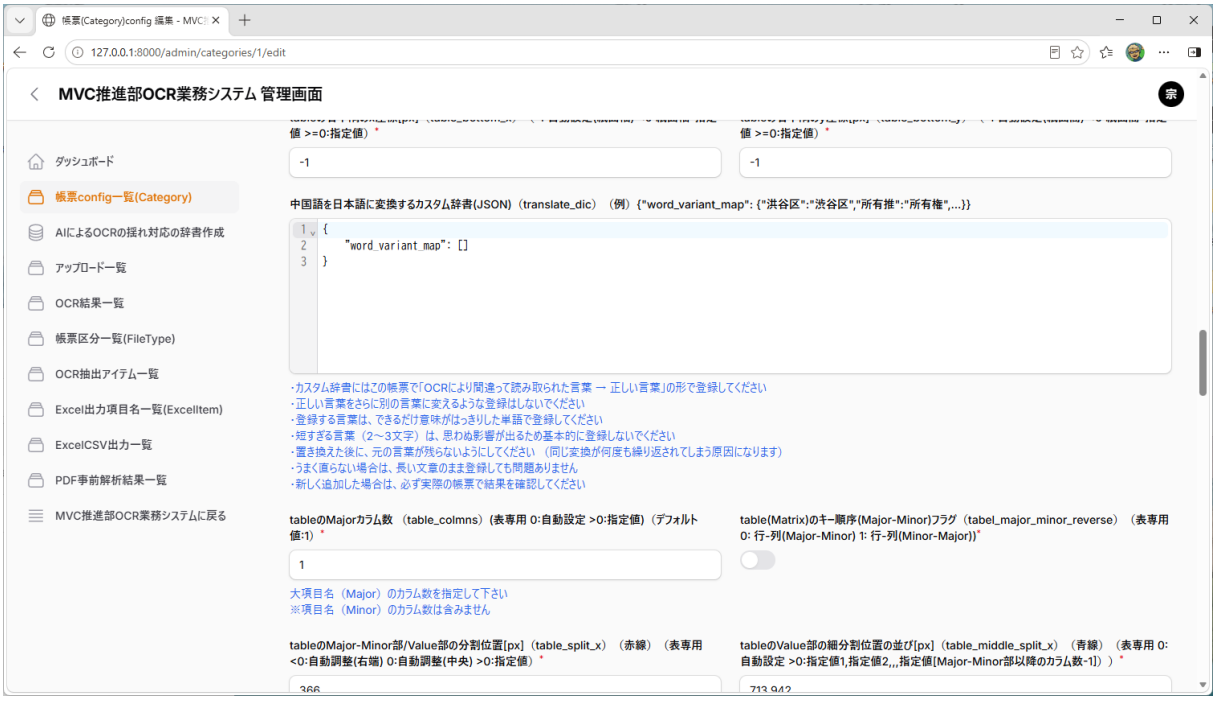
順序	操作
1	OCR 結果画面で文字揺れや誤読が残っていることを確認します。
2	「帳票(config 情報)の編集画面に進む」から帳票 config 編集画面を開き、カスタム辞書欄を確認します。
3	管理画面左メニューの「AI による OCR の揺れ対応の辞書作成」をクリックします。
4	AI による推定結果の JSON を確認し、必要に応じて編集します。登録する内容は「OCR により間違って読み取られた語句 → 正しい語句」の形にします。
5	内容を確認したうえで、「既存のカスタム辞書データとマージ保存して config 画面に戻る」をクリックします。その後、OCR 再読み込み画面に戻って OCR を再実行し、結果を確認します。

注意

カスタム辞書で登録できるのは、帳票 config 内の word_variant_map です。文字単位のシステム辞書やシステム全体の辞書は、この画面では編集しません。

日付、番号、金額など、数字部分が毎回変わる語句は、そのまま固定文字列で登録せず、必要に応じて正規表現で一般化してください。

AI の提案結果は必ず紙面と OCR 結果を確認してから保存してください。誤った辞書を登録すると、正常な文字列まで置換される場合があります。



図：帳票 config 編集画面のカスタム辞書欄



図：AI によるカスタム辞書提案画面



図：数字を含む語句の正規表現注意事項



図：カスタム辞書のマージ保存

注意

このメニューは帳票 config 編集画面を開いているときに使用します。帳票 config 編集画面以外で実行した場合は、対象となる config 情報がないためエラーになります。

7. Simple（テキスト） / 表組（TypeA） / 表組（TypeB） の使い分け

7.1 Simple（テキスト）レイアウト

Simple（テキスト）レイアウトは、赤線・青線による表の分割位置を使わず、項目名キーワードをもとに Minor / Value の組を作成する方式です。紙面上の文字の並びと OCR トークンの並びが素直な帳票では、個別の帳票設定を細かく作成しなくても、主要項目を抽出できる場合があります。

Simple（テキスト）レイアウトでは、大項目（Major）は基本的に取得されません。これは、Simple が表全体の大項目構造を解析する方式ではなく、項目名と値のペアを順番に抽出する方式であるためです。CSV 出力では主に Minor と Value を使用するため、Minor と Value が正しく取得できていれば、Major が空欄でも実用上問題ない場合があります。

Simple（テキスト）レイアウトは、表構造を厳密に固定しないため、1 つの帳票設定で複数の紙面パターンを扱いやすいという利点があります。たとえば、表だけの紙面、左側に表・右側に図面がある紙面、左側に図面・右側に表がある紙面など、OCR トークンの並びが紙面の見た目に近い帳票では、少ない設定で解析できる場合があります。

ただし、Simple（テキスト）レイアウトは万能ではありません。次のような帳票では、表組（TypeA） / 表組（TypeB） などの構造化パーサーを使用してください。

・Minor キーが行方向に結合されている表
 例：1 つの項目名が複数行の値にまたがっている表。このような行方向の結合は、表計算ソフトでは rowspan と呼ばれることがあります。本書では「行方向の結合」「連結行」と表記します。

・OCR トークンの位置や順序が紙面の見た目と大きく異なる帳票
 例：見た目では左から右、上から下に並んでいるのに、OCR 結果では別の順序で取得される帳票。

・項目名と値の対応関係が複雑な帳票
 例：値が離れた場所にある、複数の表が入り組んでいる、列方向の結合（colspan、連結カラム・連結欄）が多い帳票。

注意：Simple（テキスト）レイアウトでは、キーとして認識したい語句を帳票ごとの Simple 用キーワードリストに追加できます。また、本来は値として扱いたい語句がキーとして認識される場合は、マイナスキーワードに指定します。
システム内部であらかじめ定義されているキーワードの内容は、一般ユーザーは確認・変更できません。通常の帳票設定では、帳票ごとのキーワード追加とマイナスキーワード指定によって調整してください。

注意：キーワード登録時の空白の扱いは、Simple（テキスト）レイアウト専用ではなく帳票 config 全体の共通ルールです。Simple 用キーワードやマイナスキーワードを登録する場合も、紙面上で「備 考」「所 在 地」「価 格」のように見える語句は、「備考」「所在地」「価格」のように空白なしで登録してください。詳細は、6.2「config 編集画面の主な項目」および Appendix C のキーワード系設定の注意を参照してください。

7.2 表組（TypeA） / 表組（TypeB）

表組（TypeA） / 表組（TypeB） は、紙面の表構造を使って解析します。そのため、赤線（table_split_x）や青線（table_middle_split_x）の位置設定が重要になります。

種類	特徴
表組（TypeA）	1 行に Major / Minor / Value の組が 1 つある表組みを対象とします。
表組（TypeB）	1 行に Minor / Value の組が複数ある表組みを対象とします。赤線・青線の設定が重要です。

表組み 表組（TypeA） / 表組（TypeB） では、新規帳票で赤線・青線を確認・調整するため、通常は初回 OCR、config 調整、再 OCR という流れになります。

7.3 運用上の使い分け

帳票設定を新しく作成する場合は、まず Simple で対応できる紙面かを確認します。Simple で安定して抽出できる場合は、赤線・青線の細かい調整を行わずに運用できます。Simple で項目名と値の対応が崩れる場合は、表組（TypeA） / 表組（TypeB） の構造化パーサーを使用します。

判断基準の目安は次のとおりです。

・紙面の文字並びが素直で、OCR トークンの順序も紙面の見た目に近い
→ まず Simple（テキスト）レイアウトで確認します。

・表だけの紙面、左に表・右に図面、左に図面・右に表のような比較的単純な紙面
→ まず Simple（テキスト）レイアウトで確認します。

・本来キーになるべき項目名がキーとして認識されていない
→ その帳票の Simple 用キーワードリストに項目名を追加します。

・本来値にしたい語句がキーとして認識されている
→ その帳票のマイナスキーワードに指定します。

・備考本文が途中で分断され、通常行として出力される
→ 備考本文中でキーとして認識されている語句を、マイナスキーワードに指定します。

・Minor キーが行方向に結合されている
→ Simple ではなく表組（TypeB）などの構造化パーサーを使用します。

・1 行に 1 組の Major / Minor / Value が並ぶ
→ 表組（TypeA）を使用します。

・1 行に Minor / Value の組が複数並ぶ
→ 表組（TypeB）を使用します。

・赤線・青線を正しく指定でき、表構造を厳密に解析したい
→ 表組（TypeA） / 表組（TypeB）を使用します。

8. キーワード解析結果画面の確認

OCR 結果画面で「キーワード解析開始」ボタンをクリックすると、キーワード解析結果画面へ進みます。この画面では、OCR で抽出された項目が、基幹システムへ出力する Excel 項目にどのように対応付けられたかを確認します。


MVC推進部OCR業務システム
127.0.0.1:8000/analizer/result/119
MVC推進部OCR業務システム
管理画面
宗園巧
MVC推進部OCR業務システム
不動産・マンション情報の自動抽出ファイル（1棟収益.pdf）P.1 の 物件概要書(ミライアロー) のキーワード解析結果
物 件 概 要 書
物件名称 ミノハナビル
所在地 住所 東京都渋谷区道玄坂2丁目16-1
地番 東京都渋谷区道玄坂2-152-6
アクセス JR山手線 渋谷 駅 徒歩 3 分
東急田園都市線 渋谷 駅 徒歩 2 分
土地面積 公簿面積 29.61㎡（8.96坪）
実測面積 -
地目 地目 宅地
法規制 用途地域 商業地域
建蔽率 80%
容積率 600%
高度地区 —
日影規制 —
防火地域 防火地域
その他 ・道玄坂二丁目地区地区計画 ・渋谷駐車場整備地区
現況 賃貸中
建物 構造規模 S造 地上4階建
専有面積 90.88㎡ (27.49坪)
築年数 平成3年 1月 検査済証 -
道路 道路付け 南東側公道7.2m接道 間口約7.2m
道路計画 -
引渡 時期 相談
取引条件 相談
売却希望価格 相談 (税込)
備考
■月額賃料：1F38万円(税込) 2F～4F：40万円(税込)
■取引条件
・現況有姿 ・契約不適合責任免除
※外部への情報開示、問合せ等一切を禁じます。
発覚した際、損害賠償請求させていただきます。
二次開示資料につきましては別途CAの提出をお願いします。


MVC推進部OCR業務システム
127.0.0.1:8000/analizer/result/119
宅地建物取引業　国土交通大臣(1)第10021号
〒100−0004
東京都千代田区大手町一丁目6番1号大手町ビル Spaces大手町 320号
担当:渋谷　剛義
携帯番号:080-3565-1513 TEL:03-6841-4005 FAX:03-6841-3821
ファイル名 1棟収益.pdf
ページ数 58
SCANしたページ数 1
SCANしたページ番号 1
区分（EXCEL） 建物全部
帳票名 物件概要書(ミライアロー)
MVC推進部OCR業務システム
127.0.0.1:8000/analizer/result/119
キーワード解析結果 ※CSV出力には 項目名 (Minor) と 値 (Value) が使われます。項目区分 (Major) はOCRでは正確に取れないため参考です。
項目区分 [参考] (Major)
項目名 (Minor)
値 (Value)
確認
対応するEXCEL項目名
物件名称 物件名称 ミノハナビル → 物件名
所在地 住所 東京都渋谷区道玄坂2丁目16-1 → 市区町村以降（住居表示）
所在地 地番 東京都渋谷区道玄坂2-152-6 → 市区町村以降（地番）
アクセス JR山手線 JR山手線 渋谷 駅 徒歩3分 → 交通①
アクセス アクセス 東急田園都市線渋谷 駅 徒歩2分 → 交通②
土地面積 公簿面積 29.61㎡ (8.96坪) → 公簿地積
地目 地目 宅地 → 地目
用途地域 用途地域 商業地域 → 用途地域1
建蔽率 建蔽率 80% → 建ぺい率1
法規制 容積率 600% → 容積率1
防火地域 防火地域 防火地域 → 防火指定1



図：キーワード解析結果画面

8.1 Major / Minor / Value の考え方

キーワード解析結果画面で基本的に確認するのは、項目名（Minor）と値（Value）です。項目名は、基幹システムへ出力する Excel 項目を判定するために使用します。値は、その Excel 項目へ出力する内容です。

大項目（Major）は、項目の分類や周辺情報を示す参考情報です。CSV 出力では主に項目名と値を使用するため、大項目が空欄でも必ずしも問題ではありません。

ただし、大項目はマッピング判定の補助として使用される場合があります。たとえば、同じ「金額」という項目名でも、それが土地の金額なのか、建物の金額なのかを判断する際に、大項目が参考になることがあります。

8.2 対応する Excel 項目名の確認

自動マッピング結果は初期判定です。未設定、重複、不自然な割り当てがある場合は、Excel 出力前に必ず確認してください。

表示	意味	対応
未設定	自動判定で出力先の Excel 項目を判断できなかった状態です。	必要な情報であれば、右側の選択リストから対応する Excel 項目を手動選択します。不要な場合は未設定のままにします。
重複	同じ Excel 項目に複数の情報が割り当てられている状態です。	重複した内容は捨てられず、Excel 出力時にスペース区切りで結合されます。
丸数字付き項目	OCR から取得した複数要素の情報を一時的に受ける仮受け項目です。	Excel 出力時に必要な実項目へ展開されます。

8.3 丸数字付き仮受け項目

丸数字付きの項目は、OCR から取得した複数要素の情報を一時的に受けるための項目です。主に、交通、バス、接道、建ぺい率／容積率で使用されます。

種類	例
交通①②③	路線名、駅名、徒歩分数など。
バス①②３	バス停名、バス分数、徒歩分数など。
接道①②③④	接道方位、道路種別、幅員など。
建容①②	建ぺい率、容積率など。

同じ種類の情報が複数取得された場合は、原則として番号が順番に上がります。用意されている番号数を超える情報や、同じ Excel 項目に複数の情報が割り当てられた場合は重複として表示されますが、出力時には結合されます。

8.4 Excel 出力時の特殊処理

画面上では 1 つの項目として表示されていても、Excel / CSV へ出力する際に自動的な分解・補正処理が行われる項目があります。

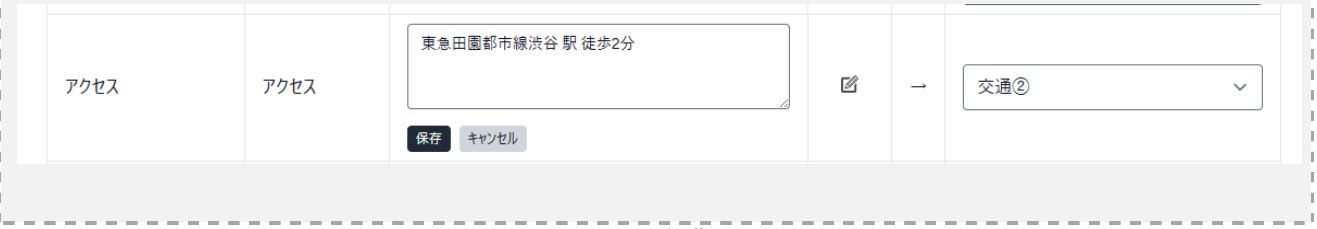
対象	処理内容
住所情報	住居表示などに都道府県名が含まれている場合、都道府県名を別フィールドへ分離して出力する場合があります。
築年月・竣工年月など	年月情報を年と月へ分解して出力する場合があります。
交通・バス・接道・建ぺい率／容積率	仮受け項目に割り当てられ、Excel / CSV 出力時に必要な実項目へ展開されます。

8.5 Value の確認と修正

キーワード解析結果画面では、OCR で抽出された内容が表形式で表示されます。OCR 結果画面のテキスト表示よりも見やすいため、OCR の誤読や値の違和感を確認しやすくなります。

値（Value）に誤りがある場合は、Value 欄の横にあるペンアイコンをクリックすると編集モードになります。編集欄に正しい内容を入力し、「保存」をクリックすると、その修正内容が Excel / CSV 出力に反映されます。

> **注意**
>
> 修正できるのは値（Value）のみです。項目名（Minor）は直接修正しません。項目名に対する出力先が違っている場合は、右側の「対応する EXCEL 項目名」の選択リストから、正しい Excel 項目を選択してください。



図：Value 編集画面

8.6 手修正と再 OCR の使い分け

状態	操作
軽微な誤読や値の一部だけの誤り	キーワード解析結果画面で Value を修正します。
対応する Excel 項目の選択だけが違う	右側の選択リストで正しい Excel 項目を選択します。
行間が狭すぎて内容が分断されている	「キーワード再解析画面に戻る」から前画面へ戻り、config を修正して再 OCR します。
本来別行の内容が巻き込まれている	config で行分割設定を修正し、OCR を再実行します。
赤線・青線が合っていない	config で table_split_x / table_middle_split_x を修正し、OCR を再実行します。

9. Excel / CSV 出力

キーワード解析結果と Excel 項目の対応を確認し、問題がなければ「EXCEL 出力」ボタンをクリックします。



図：Excel / CSV 出力ボタンと出力対象選択

出力対象	説明
すべての項目	インポート対象項目とインポート外項目の両方を出力します。
インポート項目のみ	基幹システムへ取り込む対象項目のみを出力します。通常はこちらを使用します。
インポート外項目のみ	仮受け項目や確認用項目など、基幹システムへ直接取り込まない項目を出力します。

未設定の内容は Excel / CSV には出力されません。出力したくない行は、対応する Excel 項目を「未設定」にすることで出力対象から外すことができます。

9.1 出力ファイル名

出力ファイル名には、元の PDF ファイル名、ページ番号、出力対象種別、出力日時が含まれます。ページ番号と出力日時を含めることで、同じ PDF から複数回出力した場合でもファイル名が重複しないようになっています。

PDF ファイル名_p1_excel_output_all_YYYYMMDD_HHMMSS.csv

PDF ファイル名_p1_excel_output_import_YYYYMMDD_HHMMSS.csv

PDF ファイル名_p1_excel_output_non_import_YYYYMMDD_HHMMSS.csv

例：



1 棟収益_p1_excel_output_all_20260624_101426.csv



1 棟収益_p1_excel_output_import_20260624_101414.csv



1 棟収益_p1_excel_output_non_import_20260624_101421.csv

10. 管理画面の基本操作

画面上部の「管理画面」をクリックすると、管理画面へ移動します。管理画面では、システムで使用している各種データを確認・管理できます。



図：管理画面ダッシュボード

10.1 管理画面ダッシュボードと権限

管理画面のダッシュボードには、現在登録されている所属センターの一覧が表示されます。センター情報の登録・更新は、システム管理者のみが行えます。新しいセンターの追加や既存センター情報の変更が必要な場合は、システム管理者へ依頼してください。

管理画面左側の「ユーザー一覧」以下のシステム管理系メニューは、システム管理者権限を持つユーザーにのみ表示されます。

10.2 左メニュー

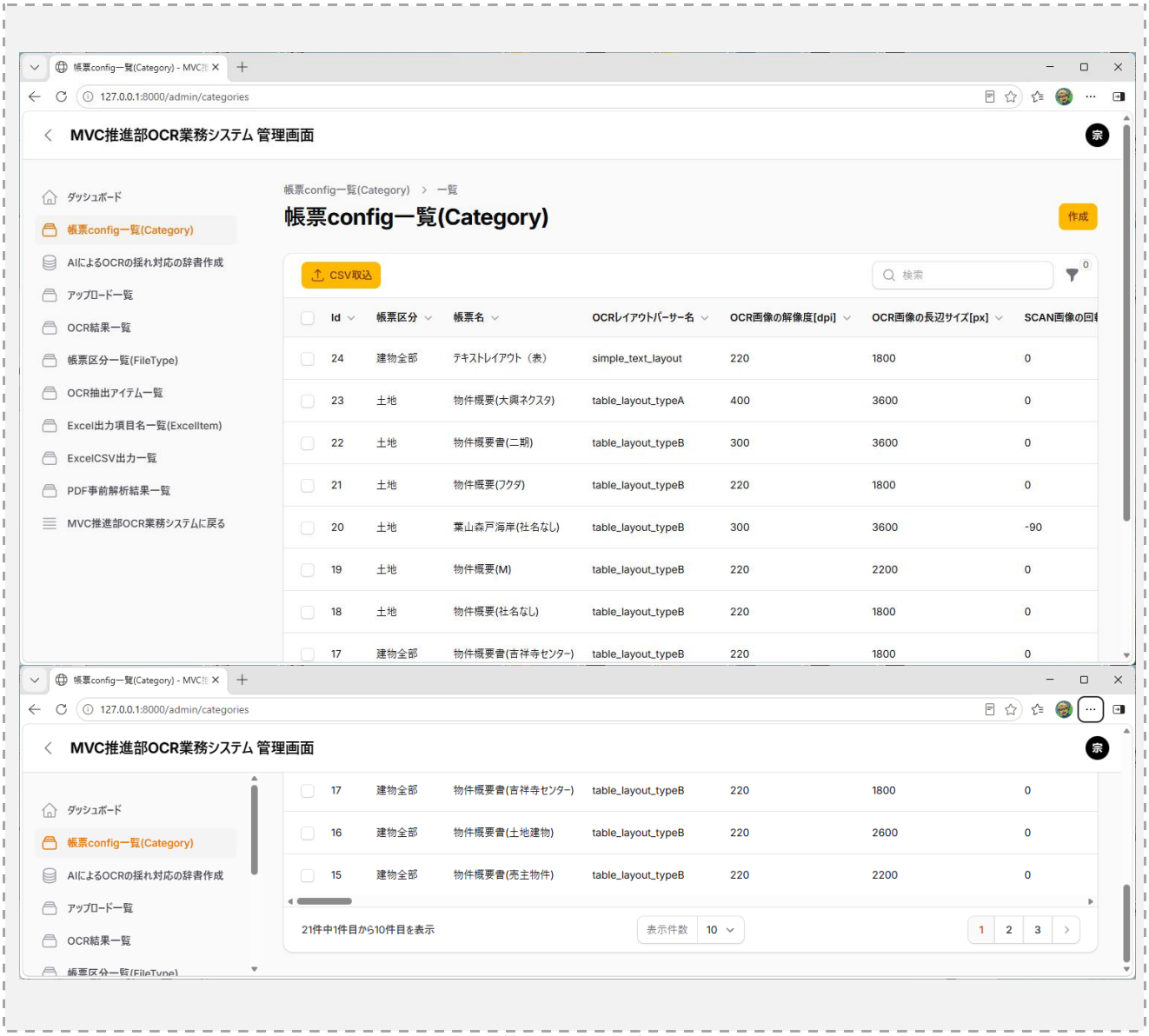
左側メニューの多くは、各データの一覧画面を表示するためのメニューです。通常利用者は、管理画面のすべての機能を使用する必要はありません。帳票設定を調整する場合は、主に「帳票 config 一覧」または OCR 結果画面から遷移する帳票 config 編集画面を使用します。

メニュー	説明
帳票 config 一覧	帳票ごとの OCR 設定を確認・編集します。
AI による OCR の揺れ対応の辞書作	帳票 config 編集画面を開いているときに使用する AI 補助機能で

メニュー	説明
成	す。
アップロード一覧	アップロードされた PDF 情報を確認します。
OCR 結果一覧	OCR 結果データを確認します。
帳票区分一覧	帳票区分情報を確認します。
OCR 抽出アイテム一覧	OCR から抽出されたアイテムを確認します。
Excel 出力項目名一覧	Excel / CSV 出力項目名を確認します。権限のある担当者のみ操作します。
ExcelCSV 出力一覧	出力結果を確認します。
MVC 推進部 OCR 業務システムに戻る	管理画面を開く前に操作していた OCR 業務システム側の画面へ戻ります。

10.3 一覧画面の共通操作

管理画面の各一覧画面は、Filament v4 の標準的な一覧画面の操作に準じています。登録済みのデータがテーブル形式で表示され、多くの一覧では ID の降順で表示されます。最近追加したデータが上に表示されるため、直近のデータを確認しやすくなっています。



図：管理画面の一覧画面

操作	説明
検索	検索欄にキーワードを入力して対象データを絞り込みます。
行クリックで編集	右端の編集ボタンまでスクロールしなくても、対象行をクリックすると編集画面へ移動できます。
作成	画面右上の「作成」ボタンから新しいデータを登録します。
CSV 取込	CSV からデータを取り込みます。
選択中の CSV 出力	チェックボックスで選択したレコードを CSV として出力します。
選択中のレコード複写	既存レコードを複写します。特に帳票 config の作成で便利です。



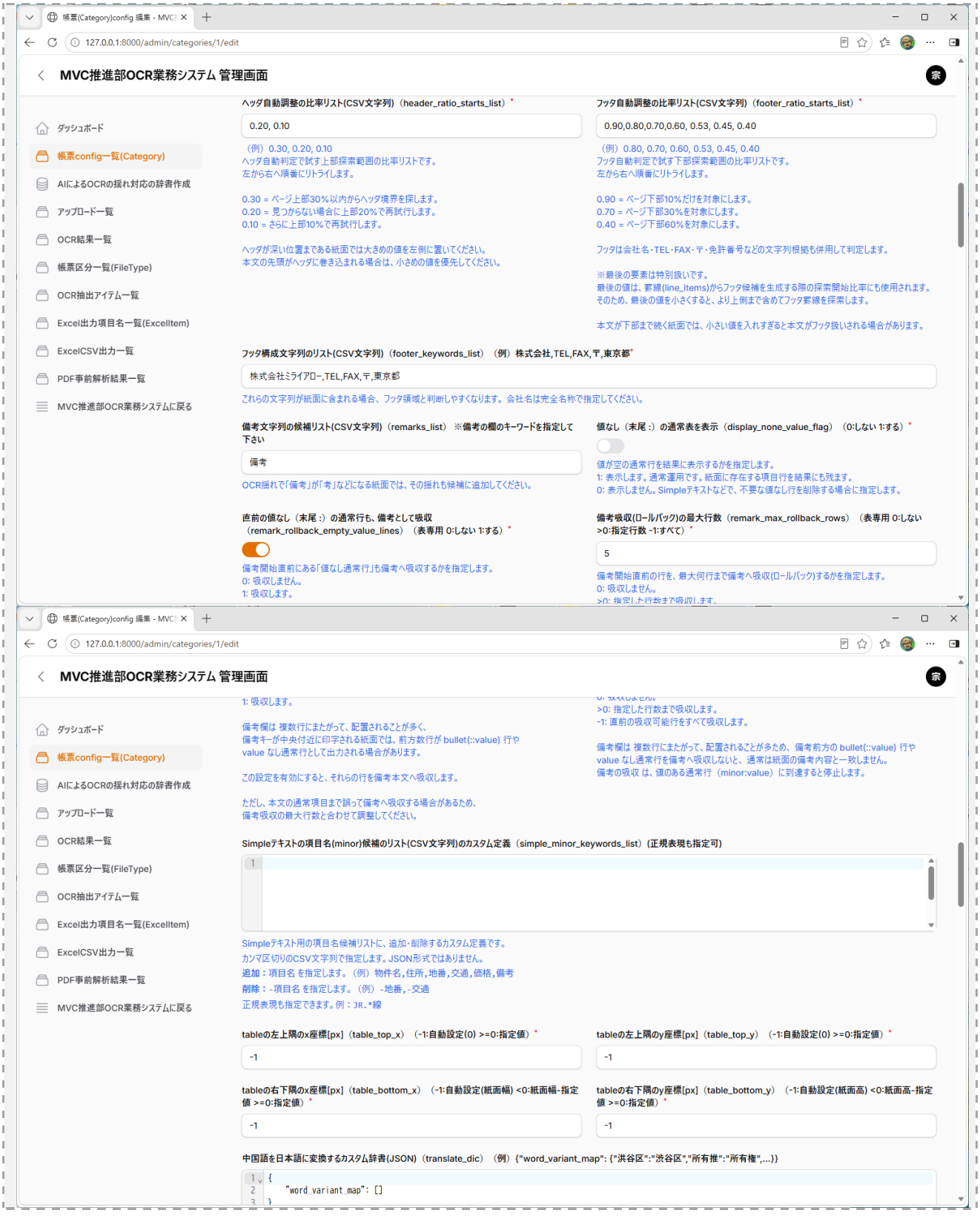
図：操作を開くメニュー

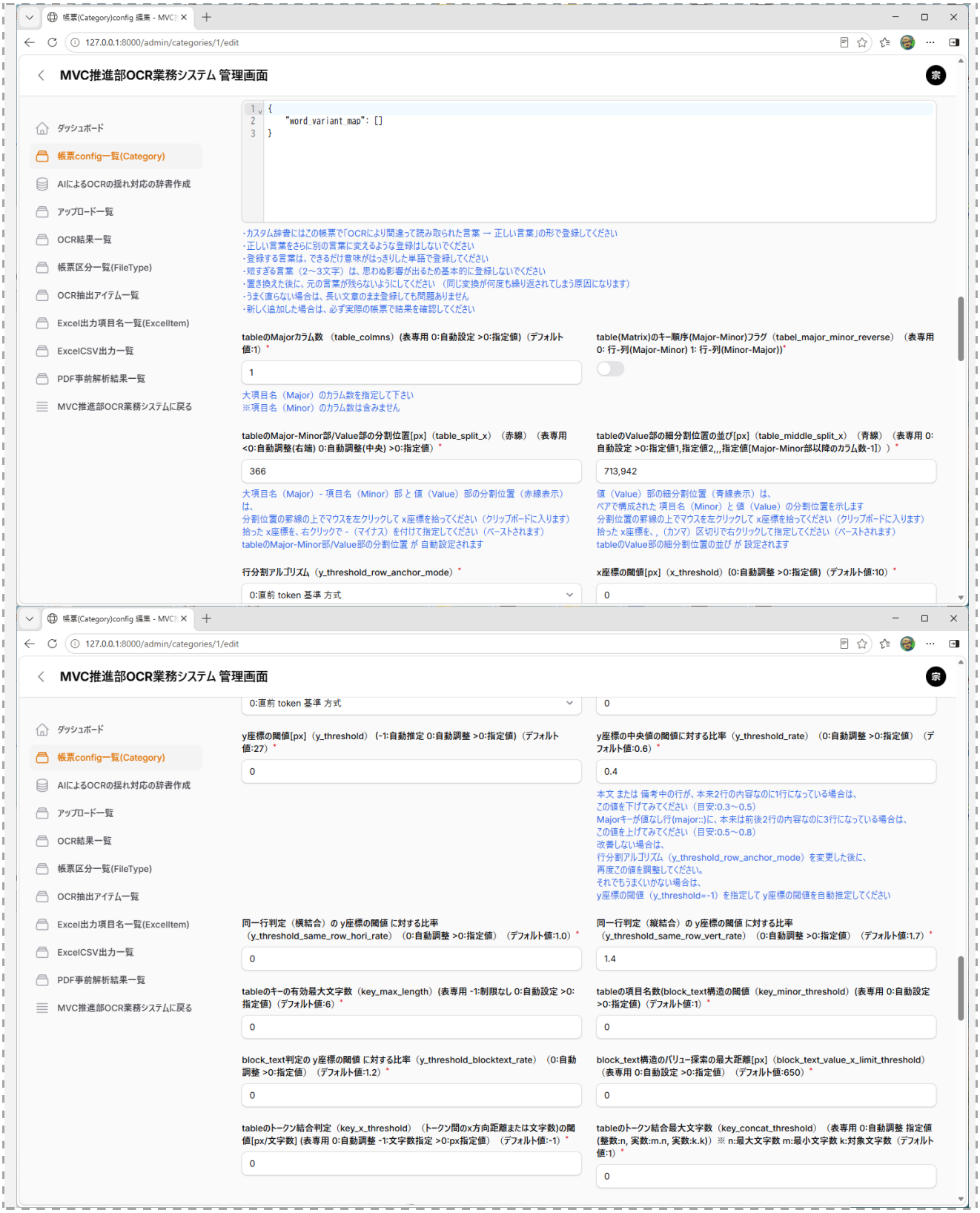
10.4 編集画面

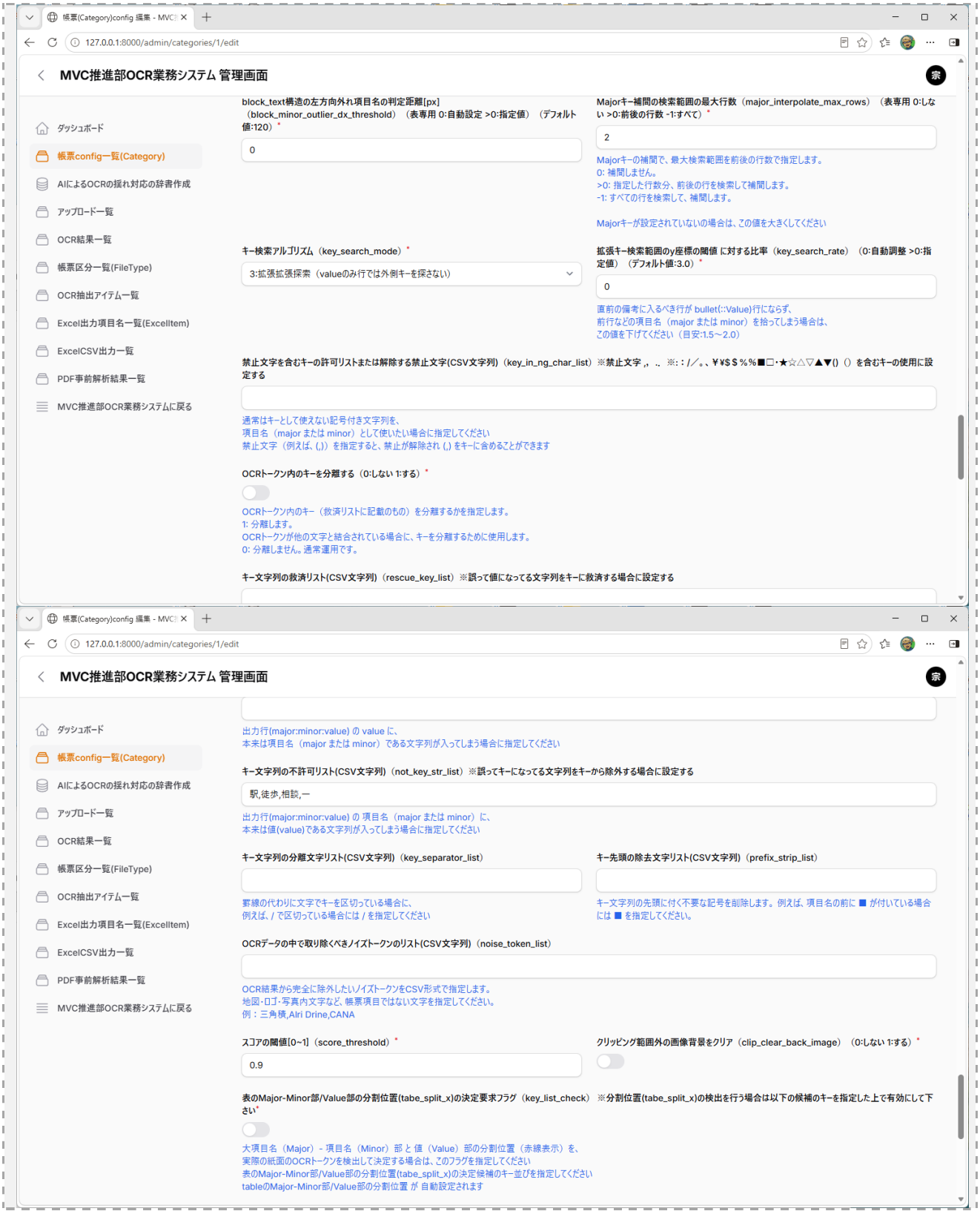
一覧画面で対象行をクリックすると、そのレコードの編集画面へ移動します。編集画面では、対象レコードに登録されている各項目を確認・修正できます。

内容を変更した場合は、画面下部の「保存」ボタンをクリックします。「キャンセル」ボタンをクリックすると、変更を保存せずに前の画面へ戻ります。保存後も同じ編集画面に残る場合があります。これは管理画面の標準動作です。


帳票(Category)config 編集 - MVC
127.0.0.1:8000/admin/categories/1/edit
MVC推進部OCR業務システム 管理画面
ダッシュボード
帳票config一覧(Category)
AIによるOCRの揺れ対応の辞書作成
アップロード一覧
OCR結果一覧
帳票区分一覧(FileType)
OCR抽出アイテム一覧
Excel出力項目名一覧(ExcelItem)
ExcelCSV出力一覧
PDF事前解析結果一覧
MVC推進部OCR業務システムに戻る
帳票config一覧(Category) > 編集
帳票(Category)config 編集
削除
サンプル帳票
x=--- y=---
table_colmns=1
click: x=--- y=---
物 件 概 要 書
物件名称
ミノハナビル
所在地
住所
東京都渋谷区道玄坂2丁目16-1
地番
東京都渋谷区道玄坂2-152-6
アクセス
JR山手線
渋谷
駅
徒歩 3 分
東急田園都市線
渋谷
駅
徒歩 2 分
土地面積
公簿面積
29.61㎡（8.96坪）
実測面積
－
地目
地目
宅地
法規制
用途地域
商業地域
建蔽率
80%
容積率
600%
高度地区
－
日影規制
帳票区分*
建物全部
帳票名*
物件概要書(ミライアロー)
OCRレイアウトパーサー名（layout）*
Table Layout Type B (Generic Pairs)
OCR画像の解像度（ocr_dpi）Simpleテキストは自動調整に戻して下さい*
220 DPI
OCR画像の長辺サイズ[px]（ocr_long_side）Simpleテキストは自動調整に戻して下さい*
1800 [px]
SCAN画像の回転[度]（ocr_rotate）Simpleテキストは自動調整に戻して下さい*
0°
config設定メモ（memo）（config調整履歴や帳票別の注意事項などの記述用です）
この帳票で使用したヘッダー・フッター,その他の調整内容、うまく動かすための注意点、変更履歴を記録します
ヘッダーの行数（header）（-1:自動判定 0:ヘッダなし >0:指定値）*
-1
ページ上部をヘッダ領域として除外する行数です。
-1: OCRトークン位置から自動判定します。
0：ヘッダなしとして扱います。
>0: 指定した行数をヘッダとして扱います。
※自動判定がうまくいかない場合に指定してください。
フッタの行数（footer）（-1:自動判定 0:フッタなし >0:指定値）*
-1
ページ下部をフッタ領域として除外する行数です。
-1: OCRトークン位置から自動判定します。
0：フッタなしとして扱います。
>0: 指定した行数をフッタとして扱います。
※自動判定がうまくいかない場合に指定してください。
ヘッダーに同じy座標の行を含める（header_line_include_flag）*
フッターに同じy座標の行を含める（footer_line_include_flag）*


帳票(Category)config 編集 - MVC
127.0.0.1:8000/admin/categories/1/edit
MVC推進部OCR業務システム 管理画面
ダッシュボード
帳票config一覧(Category)
AIによるOCRの揺れ対応の辞書作成
アップロード一覧
OCR結果一覧
帳票区分一覧(FileType)
OCR抽出アイテム一覧
Excel出力項目名一覧(ExcelItem)
ExcelCSV出力一覧
PDF事前解析結果一覧
MVC推進部OCR業務システムに戻る
ヘッダ自動調整の比率リスト(CSV文字列)（header_ratio_starts_list）*
0.20, 0.10
（例）0.30, 0.20, 0.10
ヘッダ自動判定で試す上部探索範囲の比率リストです。
左から右へ順番にリトライします。
0.30 = ページ上部30%以内からヘッダ境界を探します。
0.20 = 見つからない場合に上部20%で再試行します。
0.10 = さらに上部10%で再試行します。
ヘッダが深い位置まである紙面では大きめの値を左側に置いてください。
本文の先頭がヘッダに巻き込まれる場合は、小さめの値を優先してください。
フッタ自動調整の比率リスト(CSV文字列)（footer_ratio_starts_list）*
0.90,0.80,0.70,0.60, 0.53, 0.45, 0.40
（例）0.80, 0.70, 0.60, 0.53, 0.45, 0.40
フッタ自動判定で試す下部探索範囲の比率リストです。
左から右へ順番にリトライします。
0.90 = ページ下部10%だけを対象にします。
0.70 = ページ下部30%を対象にします。
0.40 = ページ下部60%を対象にします。
フッタは会社名・TEL・FAX・〒・免許番号などの文字列根拠も併用して判定します。
※最後の要素は特別扱いです。
最後の値は、罫線(line_items)からフッタ候補を生成する際の探索開始比率にも使用されます。
そのため、最後の値を小さくすると、より上側まで含めてフッタ罫線を探索します。
本文が下部まで続く紙面では、小さい値を入れすぎると本文がフッタ扱いされる場合があります。
フッタ構成文字列のリスト(CSV文字列)（footer_keywords_list）（例）株式会社,TEL,FAX,〒,東京都*
株式会社ミライアロー,TEL,FAX,〒,東京都
これらの文字列が紙面に含まれる場合、フッタ領域と判断しやすくなります。会社名は完全名称で指定してください。
備考文字列の候補リスト(CSV文字列)（remarks_list）※備考の欄のキーワードを指定して下さい
備考
OCR揺れで「備考」が「考」などになる紙面では、その揺れも候補に追加してください。
値なし（末尾：）の通常表を表示（display_none_value_flag）（0:しない 1:する）*
値が空の通常行を結果に表示するかを指定します。
1: 表示します。通常運用です。紙面に存在する項目行を結果にも残ます。
0: 表示しません。Simpleテキストなどで、不要な値なし行を削除する場合に指定します。
直前の値なし（末尾：）の通常行も、備考として吸収（remark_rollback_empty_value_lines）（表専用 0:しない 1:する）*
備考開始直前にある「値なし通常行」も備考へ吸収するかを指定します。
0: 吸収しません。
1: 吸収します。
備考吸収(ロールバック)の最大行数（remark_max_rollback_rows）（表専用 0:しない >0:指定行数 -1:すべて）*
5
備考開始直前の行を、最大何行まで備考へ吸収(ロールバック)するかを指定します。
0: 吸収しません。
>0: 指定した行数まで吸収します。
-1: 直前の吸収可能行をすべて吸収します。
備考欄は 複数行にまたがって、配置されることが多く、
備考キーが中央付近に印字される紙面では、前方数行が bullet(::value) 行や
value なし通常行として出力される場合があります。
この設定を有効にすると、それらの行を備考本文へ吸収します。
ただし、本文の通常項目まで誤って備考へ吸収する場合があるため、
備考吸収の最大行数と合わせて調整してください。
備考欄は 複数行にまたがって、配置されることが多ため、備考前方の bullet(::value) 行や
value なし通常行を備考へ吸収しないと、通常は紙面の備考内容と一致しません。
備考の吸収 は、値のある通常行（minor:value）に到達すると停止します。
Simpleテキストの項目名(minor)候補のリスト(CSV文字列)のカスタム定義（simple_minor_keywords_list）(正規表現も指定可)
Simpleテキスト用の項目名候補リストに、追加・削除するカスタム定義です。
カンマ区切りのCSV文字列で指定します。JSON形式ではありません。
追加：項目名 を指定します。（例）物件名,住所,地番,交通,価格,備考
削除：-項目名 を指定します。（例）-地番,-交通
正規表現も指定できます。例：JR.*線
tableの左上隅のx座標[px]（table_top_x）（-1:自動設定(0) >=0:指定値）*
-1
tableの左上隅のy座標[px]（table_top_y）（-1:自動設定(0) >=0:指定値）*
-1
tableの右下隅のx座標[px]（table_bottom_x）（-1:自動設定(紙面幅) <0:紙面幅-指定値 >=0:指定値）*
-1
tableの右下隅のy座標[px]（table_bottom_y）（-1:自動設定(紙面高) <0:紙面高-指定値 >=0:指定値）*
-1
中国語を日本語に変換するカスタム辞書(JSON)（translate_dic）（例）{"word_variant_map": {"渋谷区":"渋谷区","所有推":"所有権",...}}
{
"word_variant_map": []
}


帳票(Category)config 編集 - MVC
127.0.0.1:8000/admin/categories/1/edit
MVC推進部OCR業務システム 管理画面
ダッシュボード
帳票config一覧(Category)
AIによるOCRの揺れ対応の辞書作成
アップロード一覧
OCR結果一覧
帳票区分一覧(FileType)
OCR抽出アイテム一覧
Excel出力項目名一覧(ExcelItem)
ExcelCSV出力一覧
PDF事前解析結果一覧
MVC推進部OCR業務システムに戻る
{
"word_variant_map": []
}
・カスタム辞書にはこの帳票で「OCRにより間違って読み取られた言葉 → 正しい言葉」の形で登録してください
・正しい言葉をさらに別の言葉に変えるような登録はしないでください
・登録する言葉は、できるだけ意味がはっきりした単語で登録してください
・短すぎる言葉（2〜3文字）は、思わぬ影響が出るため基本的に登録しないでください
・置き換えた後に、元の言葉が残らないようにしてください（同じ変換が何度も繰り返されてしまう原因になります）
・うまく直らない場合は、長い文章のまま登録しても問題ありません
・新しく追加した場合は、必ず実際の帳票で結果を確認してください
tableのMajorカラム数（table_colmns）(表専用 0:自動設定 >0:指定値)（デフォルト値:1）*
1
大項目名（Major）のカラム数を指定して下さい
※項目名（Minor）のカラム数は含みません
table(Matrix)のキー順序(Major-Minor)フラグ（tabel_major_minor_reverse）（表専用 0: 行-列(Major-Minor) 1: 行-列(Minor-Major)）*
tableのMajor-Minor部/Value部の分割位置[px]（table_split_x）（赤線）（表専用 <0:自動調整(右端) 0:自動調整(中央) >0:指定値）*
366
大項目名（Major）- 項目名（Minor）部 と 値（Value）部の分割位置（赤線表示）は、
分割位置の罫線の上でマウスを左クリックして x座標を拾ってください（クリップボードに入ります）
拾った x座標を、右クリックで - （マイナス）を付けて指定してください（ペーストされます）
tableのMajor-Minor部/Value部の分割位置 が 自動設定されます
tableのValue部の細分割位置の並び[px]（table_middle_split_x）（青線）（表専用 0:自動設定 >0:指定値1,指定値2,,,指定値[Major-Minor部以降のカラム数-1]）*
713,942
値（Value）部の細分割位置（青線表示）は、
ペアで構成された 項目名（Minor）と 値（Value）の分割位置を示します
分割位置の罫線の上でマウスを左クリックして x座標を拾ってください（クリップボードに入ります）
拾った x座標を、,（カンマ）区切りで右クリックして指定してください（ペーストされます）
tableのValue部の細分割位置の並び が 設定されます
行分割アルゴリズム（y_threshold_row_anchor_mode）*
0:直前 token 基準 方式
x座標の閾値[px]（x_threshold）(0:自動調整 >0:指定値)（デフォルト値:10）*
0
y座標の閾値[px]（y_threshold）(-1:自動推定 0:自動調整 >0:指定値)（デフォルト値:27）*
0
y座標の中央値の閾値に対する比率（y_threshold_rate）（0:自動調整 >0:指定値）（デフォルト値:0.6）*
0.4
本文 または 備考中の行が、本来2行の内容なのに1行になっている場合は、
この値を下げてみてください（目安:0.3〜0.5）
Majorキーが値なし行(major::)に、本来は前後2行の内容なのに3行になっている場合は、
この値を上げてみてください（目安:0.5〜0.8）
改善しない場合は、
行分割アルゴリズム（y_threshold_row_anchor_mode）を変更した後に、
再度この値を調整してください。
それでもうまくいかない場合は、
y座標の閾値（y_threshold=-1）を指定して y座標の閾値を自動推定してください
同一行判定（横結合）の y座標の閾値 に対する比率（y_threshold_same_row_hori_rate）（0:自動調整 >0:指定値）（デフォルト値:1.0）*
0
同一行判定（縦結合）の y座標の閾値 に対する比率（y_threshold_same_row_vert_rate）（0:自動調整 >0:指定値）（デフォルト値:1.7）*
1.4
tableのキーの有効最大文字数（key_max_length）(表専用 -1:制限なし 0:自動設定 >0:指定値)（デフォルト値:6）*
0
tableの項目名数(block_text構造の閾値（key_minor_threshold）(表専用 0:自動設定 >0:指定値)（デフォルト値:1）*
0
block_text判定の y座標の閾値 に対する比率（y_threshold_blocktext_rate）（0:自動調整 >0:指定値）（デフォルト値:1.2）*
0
block_text構造のバリュー探索の最大距離[px]（block_text_value_x_limit_threshold）（表専用 0:自動設定 >0:指定値）（デフォルト値:650）*
0
tableのトークン結合判定（key_x_threshold）（トークン間のx方向距離または文字数)の閾値[px/文字数] (表専用 0:自動調整 -1:文字数指定 >0:px指定値）（デフォルト値:-1）*
0
tableのトークン結合最大文字数（key_concat_threshold）（表専用 0:自動調整 指定値(整数:n, 実数:m.n, 実数:k.k)）※ n:最大文字数 m:最小文字数 k:対象文字数（デフォルト値:1）*
0


帳票(Category)config 編集 - MVC
127.0.0.1:8000/admin/categories/1/edit
MVC推進部OCR業務システム 管理画面
家
ダッシュボード
帳票config一覧(Category)
AIによるOCRの揺れ対応の辞書作成
アップロード一覧
OCR結果一覧
帳票区分一覧(FileType)
OCR抽出アイテム一覧
Excel出力項目名一覧(ExcelItem)
ExcelCSV出力一覧
PDF事前解析結果一覧
MVC推進部OCR業務システムに戻る
block_text構造の左方向外れ項目名の判定距離[px]（block_minor_outlier_dx_threshold）（表専用 0:自動設定 >0:指定値）（デフォルト値:120）*
0
Majorキー補間の検索範囲の最大行数（major_interpolate_max_rows）（表専用 0:しない >0:前後の行数 -1:すべて）*
2
Majorキーの補間で、最大検索範囲を前後の行数で指定します。
0: 補間しません。
>0: 指定した行数分、前後の行を検索して補間します。
-1: すべての行を検索して、補間します。
Majorキーが設定されていないの場合は、この値を大きくしてください
キー検索アルゴリズム（key_search_mode）*
3:拡張拡張探索（valueのみ行では外側キーを探さない）
拡張キー検索範囲のy座標の閾値 に対する比率（key_search_rate）（0:自動調整 >0:指定値）（デフォルト値:3.0）*
0
直前の備考に入るべき行が bullet(::Value)行にならず、
前行などの項目名（major または minor）を拾ってしまう場合は、
この値を下げてください（目安:1.5～2.0）
禁止文字を含むキーの許可リストまたは解除する禁止文字(CSV文字列)（key_in_ng_char_list）※禁止文字 ,, .. ※: : / ／ 。、￥¥$ $ % %■□・★☆△▽▲▼() （）を含むキーの使用に設定する
通常はキーとして使えない記号付き文字列を、
項目名（major または minor）として使いたい場合に指定してください
禁止文字（例えば、(,)）を指定すると、禁止が解除され (,) をキーに含めることができます
OCRトークン内のキーを分離する（0:しない 1:する）*
OCRトークン内のキー（救済リストに記載のもの）を分離するかを指定します。
1: 分離します。
OCRトークンが他の文字と結合されている場合に、キーを分離するために使用します。
0: 分離しません。通常運用です。
キー文字列の救済リスト(CSV文字列)（rescue_key_list）※誤って値になってる文字列をキーに救済する場合に設定する
出力行(major:minor:value) の value に、
本来は項目名（major または minor）である文字列が入ってしまう場合に指定してください
キー文字列の不許可リスト(CSV文字列)（not_key_str_list）※誤ってキーになってる文字列をキーから除外する場合に設定する
駅,徒歩,相談,一
出力行(major:minor:value) の 項目名（major または minor）に、
本来は値(value)である文字列が入ってしまう場合に指定してください
キー文字列の分離文字リスト(CSV文字列)（key_separator_list）
罫線の代わりに文字でキーを区切っている場合に、
例えば、/ で区切っている場合には / を指定してください
キー先頭の除去文字リスト(CSV文字列)（prefix_strip_list）
キー文字列の先頭に付く不要な記号を削除します。例えば、項目名の前に ■ が付いている場合には ■ を指定してください。
OCRデータの中で取り除くべきノイズトークンのリスト(CSV文字列)（noise_token_list）
OCR結果から完全に除外したいノイズトークンをCSV形式で指定します。
地図・ロゴ・写真内文字など、帳票項目ではない文字を指定してください。
例：三角積,Alri Drine,CANA
スコアの閾値[0~1]（score_threshold）*
0.9
クリッピング範囲外の画像背景をクリア（clip_clear_back_image）（0:しない 1:する）*
表のMajor-Minor部/Value部の分割位置(tabe_split_x)の決定要求フラグ（key_list_check）※分割位置(tabe_split_x)の検出を行う場合は以下の候補のキーを指定した上で有効にして下さい*
大項目名（Major）- 項目名（Minor）部 と 値（Value）部の分割位置（赤線表示）を、
実際の紙面のOCRトークンを検出して決定する場合は、このフラグを指定してください
表のMajor-Minor部/Value部の分割位置(tabe_split_x)の決定候補のキー並びを指定してください
tableのMajor-Minor部/Value部の分割位置 が 自動設定されます



図：管理画面の編集画面

10.5 管理画面右上のユーザーメニュー

管理画面右上のユーザーアイコンをクリックすると、ユーザーメニューが表示されます。このメニューからログアウトできます。また、画面表示のテーマ切替など、Filament 標準の表示設定が表示される場合があります。通常の OCR 操作では、これらの表示設定を変更する必要はありません。

本システムでは、OCR 業務システム側と管理画面側でログイン情報が共有されています。OCR 業務システムにログインした状態で管理画面へ移動すると、再ログインせずにそのまま管理画面を利用できます。



図：管理画面右上のユーザーメニュー

11. 困ったときの確認ポイント

現象	確認する画面・設定	対応
文字が誤っている	OCR 結果画面、整形前テキスト、translate_dic	Value を手修正するか、translate_dic へ辞書登録します。
同じ住所などが複数行に分かれる	行分割データ、y_threshold_rate	y_threshold_rate を大きくして再 OCR します。
別行が同じ行に巻き込まれる	行分割データ、y_threshold_rate	y_threshold_rate を小さくして再 OCR します。
同じトークンが複数行に重複する	行分割データ、y_threshold_rate	y_threshold_rate を小さくして再 OCR します。
赤線・青線がずれている	画像プレビュー、table_split_x、table_middle_split_x	画像をクリックして X 座標を取得し、設定へ貼り付けます。
地図・写真・ロゴが混入する	クリッピング範囲	table_top_x/y、table_bottom_x/y で表領域を限定します。
未設定が多い	キーワード解析結果画面、Excel 項目選択	必要な項目は選択リストから手動で Excel 項目を選択します。
軽微な Value 誤り	キーワード解析結果画面	ペンアイコンから Value を修正します。
手修正で対応できないほど崩れている	キーワード再解析画面、帳票 config 編集	キーワード再解析画面に戻り、config を修正して再 OCR します。

Appendix A. OCR パーサーの種類と表構造について

本付録以降は、通常利用者向けの操作説明ではなく、帳票を新規作成・調整する担当者、システム担当者、保守担当者向けの技術資料です。OCR 結果が崩れた場合に、どのパーサーを選ぶべきか、どの config を調整すべきか、正規化データをどう読むべきかを判断するための前提知識として使用してください。

OCR 処理では、紙面固有の即値やリテラルに依存せず、OCR トークンの bbox、相対位置、score、config 値を根拠として、紙面を共通形式へ変換します。

A.1 OCR パーサーの役割

OCR エンジンは、紙面上の文字を位置情報付きのトークンとして読み取ります。OCR パーサーは、そのトークンを行・列・表領域・項目名・値の関係に整理し、最終的に major / minor / value の正規化データへ変換します。

パーサーの選択と config 調整を誤ると、文字自体は読めていても、項目名と値の対応関係が崩れます。そのため、帳票作成・調整時には、対象紙面が Simple / 表組（TypeA） / 表組（TypeB） のどの構造に近いかを先に判断します。

A.2 表の配置範囲

表が紙面全体に配置される場合と、紙面の一部だけに配置される場合があります。地図・写真・ロゴ・余白文字などが混在する場合は、table_top_x / table_top_y / table_bottom_x / table_bottom_y で処理対象の矩形範囲を指定します。

[Type ①：単一紙面での表の配置]



表が紙面全体に配置される場合:

x0=0, y0=0, x1=紙面の幅[px], y1=紙面の高さ[px]

[Type ②：分割紙面での表の配置]



表だけを対象にする場合:

`table_top_x / table_top_y / table_bottom_x / table_bottom_y` で矩形範囲を指定する。

A.3 Simple（テキスト）レイアウト

Simple（テキスト）レイアウトは、表の罫線や赤線・青線による分割位置を厳密には使用しません。simple_minor_keywords_list に登録された項目名候補を使い、見つかった項目名から次の項目名までを値として扱います。

Simple では Major は原則として使用しません。そのため、出力は主に「:minor:value」または「::value」の形式になります。OCR 結果としては構造情報が少なくなりますが、minor と value が正しく取得できれば、CSV マッピングで十分に救済できる場合があります。

[Simple：項目名キーワードによる抽出]

```
+-------------------------------------------------------------+
| xxx xxx  Keyword1  yyy yyy ... KeywordN  zzz zzz zzz      |
+-------------------------------------------------------------+
```

simple_minor_keywords_list = Keyword1, Keyword2, ...

出力例:

```
 ::xxx xxx
 :Keyword1:yyy yyy
 :KeywordN:zzz zzz zzz
```

特徴:

- 赤線/青線には依存しない
- Major は原則として生成しない
- 拾えない文字列は ::value に逃がし、構造破壊を避ける

Simple（テキスト）レイアウトの考え方

Simple（テキスト）レイアウトは、OCR トークンを紙面上の位置に従って並べ、Simple 用キーワードとして認識した語句を Minor として扱います。そして、その Minor から次の Minor までの間にある文字列を Value として扱います。

例：

物件名称　品川区大井 3 丁目土地　交通　JR 京浜東北線「大井町」駅 徒歩 10 分

出力例：
:物件名称:品川区大井 3 丁目土地
:交通:JR 京浜東北線「大井町」駅 徒歩 10 分

このため、Simple（テキスト）レイアウトは表組（TypeB）に近い考え方で、キーと値のペアを順番に抽出します。ただし、表組（TypeB）のように赤線・青線で表の列構造を指定するわけではありません。表構造を厳密に見ない代わりに、項目名キーワードと OCR トークンの並びをもとに解析します。

Simple（テキスト）レイアウトが向いている帳票

Simple（テキスト）レイアウトは、次の条件を満たす帳票に向いています。
・OCR トークンの並びが紙面の見た目に近い
・Minor キーと Value が紙面上で明確なペアになっている
・Minor キーが行方向に結合されていない
・表だけ、または表と図面が左右に分かれているなど、配置が比較的素直である
・帳票ごとに赤線・青線を細かく調整するよりも、まず主要項目を効率よく抽出したい

Simple（テキスト）レイアウトが向いていない帳票

次のような帳票では、Simple では正しい Minor / Value の対応を取れない場合があります。

・Minor キーが行方向に結合されている表
1 つの項目名が複数行の値にまたがる表です。表計算ソフトでは rowspan と呼ばれることがあります。本書では「行方向の結合」「連結行」と表記します。

・OCR トークンの並びが紙面の見た目と大きく異なる帳票
古い PDF や特殊な帳票では、見た目の位置と無関係な順序で OCR トークンが取得されることがあります。この場合、Simple は項目名と値の順序を正しく判断できない可能性があります。

・列方向の結合が多い表
複数列にまたがる欄です。表計算ソフトでは colspan と呼ばれることがあります。本書では「列方向の結合」「連結カラム」「連結欄」と表記します。

・値が離れた場所にある表、複数の表が入り組んでいる帳票
この場合は、表領域や列構造を指定できる表組（TypeA） / 表組（TypeB）を使用します。

Simple（テキスト）レイアウトで調整できること

Simple（テキスト）レイアウトで利用者が調整できる主な内容は、次の 2 つです。
1. キーとして認識したい語句を、その帳票の Simple 用キーワードリストに追加する
2. 本来は値として扱いたい語句がキーとして認識される場合、その帳票のマイナスキーワードに指定する

システム内部であらかじめ定義されているキーワードの内容は、一般ユーザーは確認・変更できません。帳票ごとの設定では、Simple 用キーワードリストへの追加と、マイナスキーワード指定によって調整します。

注意：キーワード登録時の空白の扱いは、Simple（テキスト）レイアウト専用ではなく帳票 config 全体の共通ルールです。Simple 用キーワードリストやマイナスキーワードでも、紙面上で「備 考」「所 在地」のように文字間に空白がある場合は、「備考」「所在地」のように空白なしで登録してください。

備考ブロックでの注意

備考本文中の語句がキーとして認識されると、備考ブロックが途中で分断され、本来は備考本文として連結されるべき内容が通常行として出力される場合があります。

この場合は、通常行として出てしまった語句を確認し、その帳票のマイナスキーワードに指定してください。これにより、その語句はその帳票ではキーとして扱われず、備考本文または値側の文字列として扱われます。

A.4 表組（TypeA）レイアウト

表組（TypeA）は、1 行に Major / Minor / Value の組が 1 つある表です。内容（Value）の左側に、Major カラムと Minor カラムがすべて並ぶ構造を想定します。

table_colmns は Major カラム数を指定します。Major カラムが複数ある場合、出力される major は各 Major 要素をハイフンで連結した文字列になります。たとえば major1=土地、major2=面積 の場合、major は「土地-面積」として扱います。

表組（TypeA）の赤線（table_split_x）は、Major/Minor 部と Value 部の境界です。

[表組（TypeA）：一般系 N 段組み]

```
+-----------------------------------------------+----------------+
| major1 major2 ... majorN  minor               | value          |
|                                               ↑                |
|                                          table_split_x         |
+-----------------------------------------------+----------------+
```

出力形式:

[major]:[minor]:value

Major が複数ある場合:

major = major1 + '-' + major2 + '-' + ... + '-' + majorN

A.5 表組（TypeB）レイアウト

表組（TypeB）は、1 行の中に Minor / Value のペアが複数並ぶ表です。左側に Major カラム群があり、右側に複数の Minor / Value ペアが並ぶ構造を想定します。

表組（TypeB）では、赤線（table_split_x）で Major 部と右側のペア領域を分け、青線（table_middle_split_x）で Value 部の中にある追加分割位置を指定します。青線がずれると、右側の項目名と値の組が崩れます。

Major カラムが複数ある場合、表組（TypeA）と同様に、出力される major は各 Major 要素をハイフンで連結した文字列になります。

[表組（TypeB）：Minor / Value ペアが横に複数並ぶ表]

```
+-------------------+-------------------------------------------------------------+
| major1 major2 ... | minor1 value1 | minor2 value2 | minor3 v3                   |
|                                        ↑                ↑             ↑       |
|                                  table_split_x        ^blue1        blue2      |
+-------------------+-------------------------------------------------------------+
```

処理イメージ：

major = major1 + '-' + major2 + ... + '-' + majorN

各ペアを順に処理し、以下の行を出力する。

major:minor1:value1

major:minor2:value2

major:minor3:value3

青線 table_middle_split_x:

Value 部内の追加分割位置をカンマ区切りで指定する。

A.6 赤線・青線の指定方法

赤線（table_split_x）は、Major/Minor 部と Value 部の境界です。赤線は、罫線そのものの位置ではなく、OCR トークンの中心座標（cx）を基準に分割位置を計算します。そのため、config で固定指定する場合は、クリックした x 座標をマイナス付きの値として指定します。例：クリック位置が 1807 の場合は -1807 を指定します。

青線（table_middle_split_x）は、Value 部の中の追加分割位置です。青線は、画面左端を基準とした x 座標をそのまま使用します。そのため、赤線のようにマイナスにする必要はありません。複数指定する場合は、713,942 のようにカンマ区切りで指定します。

クリック座標は、帳票 config 編集画面の OCR 画像上で取得できます。取得した座標を config 項目へ貼り付けた後、OCR を再実行し、「OCR パーサーの抽出された整形前テキスト」で分割結果を確認します。

A.7 レイアウト選択の目安

紙面上の文字の並びが素直で、項目名キーワードから値を拾える場合は Simple を使用します。1 行に 1 組の Major / Minor / Value が並ぶ表は 表組（TypeA） を使用します。1 行に Minor / Value のペアが複数並ぶ表は 表組（TypeB） を使用します。表領域が紙面の一部に限られる場合は、パーサー種別に関係なく、表の配置範囲を先に指定します。

判断に迷う場合は、まず OCR 結果画面の画像上で表の構造を確認し、次に「OCR パーサーの行分割データ」と「OCR パーサーの抽出された整形前テキスト」で、行・列・key/value の分かれ方を確認します

Appendix B. OCR パーサーが出力する正規化データについて

本付録では、OCR パーサーが出力する正規化データの形式を、技術的・文法的な観点から説明します。帳票 config 調整、OCR 結果の原因調査、キーワード解析結果の確認、CSV 出力不具合の切り分けを行う担当者向けの内容です。

B.1 正規化データの目的

OCR エンジンは、紙面上の文字を位置情報付きのトークンとして読み取ります。OCR パーサーは、その結果を行単位・項目単位に整理し、後段の Laravel、CSV 出力、Salesforce 連携が扱いやすいように、必ず同じ形の行データへ正規化します。

全体の目的は、人が確認・補正でき、ルールベースで同じ構造に正規化でき、後段処理を単純にできる状態にすることです。

B.2 正規化された行出力の基本形式

OCR パーサの正規化された行出力の形式

[[？]major]:[？]minor:[？]value　？は低信頼マーク　[] は省略可能を示す記号

例外は 備考 の出力の形式

備考:備考: ここは改行コードを含む備考内容（複数行） になります

CSV 出力に利用するのは `minor:value` のみである

major　人が確認するためのもので、項目区分を表す（OCR は行単位の SCAN のためほとんどとれない）

minor　項目名の候補（EXCEL の本当の公務区とマッピングするための情報）

value　項目の値 （※備考の場合は 備考内容（複数行も可））

B.3 行タイプと省略形

正規化データでは、半角コロン「:」で major / minor / value を区切ります。空欄がある場合も、位置関係が崩れないようにコロンを残します。

通常行	`major:minor:value`
`major` のみ	`major::value`
`minor` のみ	`:minor:value`
`minor` のみ・`value` なし	`:minor:`
`bullet` 行	`::value`
備考	備考:備考:`value`（複数行可）
ヘッダ	ヘッダ`::value`
フッタ	フッタ`::value`

minor の位置は CSV マッピングで重要です。major が空欄でも、minor と value が正しく入っていれば、CSV 出力では問題にならない場合があります。

B.4 低信頼マーク「？」の扱い

OCR score が score_threshold を下回る要素を含む場合、その要素の先頭に「？」が付くことがあります。これは利用者・調整担当者に確認を促すためのマークです。

低信頼マークは要素の先頭に 1 つだけ付与されます。キーワード解析時には自動的に除去されますが、元の文字が誤っている場合は、Value 修正または translate_dic への辞書登録で対応します。

B.5 major / minor / value の意味

major は、土地、建物、法規制、設備などの項目区分を表します。ただし、OCR は行単位で処理するため、表の大項目が複数行にまたがる場合などは、すべての行で major を正確に取得できるとは限りません。

minor は、所在地、地目、建ぺい率、価格など、Excel 項目とマッピングするための項目名候補です。CSV 出力では、基本的に minor と value が重要になります。

value は、minor に対応する実際の値です。備考の場合は、複数行の文章を改行付きで保持することがあります。

B.6 備考・備考モードの扱い

備考欄は、通常の表項目とは異なり、複数行の文章や箇条書きをまとめて保持する必要があります。そのため、備考候補文字列（remarks_list）に一致した行を起点として、備考モードで周辺行を吸収します。

備考モードの考え方:

開始条件:

major または minor が remarks_list の候補に一致する。

モード中の主な扱い:

・備考直前の bullet 行（::value）は pre_remark_buffer に退避する。

・備考行に到達したら、退避分と備考自身を結合する。

・必要に応じて、直前の value なし通常行も備考へ吸収する。

・フッタ検出または次の通常項目検出で備考モードを終了する。

最終出力:

備考:備考:複数行の備考内容

備考欄では、文章を細かい項目に分解しすぎると情報が失われるため、元の内容を保持することを優先します。備考として吸収する行数は remark_max_rollback_rows、value なし行を吸収するかどうかは remark_rollback_empty_value_lines で調整します。

B.7 ヘッダ・フッタの扱い

ヘッダ行・フッタ行は、CSV 出力対象ではなく、紙面上の説明・会社情報・注記などとして扱います。ヘッダ・フッタの判定には、行位置だけでなく footer_keywords_list などの内容的根拠も使用します。

フッタが本文や備考に混入すると、後続のマッピングや備考吸収に影響します。備考との境界が崩れる場合は、footer、footer_ratio_starts_list、footer_keywords_list を確認します。

B.8 Simple（テキスト）レイアウトの場合

Simple（テキスト）レイアウトでは、Major は原則として生成しません。項目名キーワードを minor として扱い、その右側または次のキーワードまでの文字列を value として出力します。項目名として拾えない文字列は、無理に major/minor を作らず、::value として保持します。

この設計により、Simple（テキスト）レイアウトは表構造の情報量は少ないものの、構造が壊れにくく、CSV マッピングで救済しやすい正規化データを出力できます。

Simple（テキスト）レイアウトでは、Minor / Value の対応が正しく取れているかを確認します。Major が空欄であっても、Minor と Value が正しく取得できていれば、CSV マッピングでは問題ない場合があります。

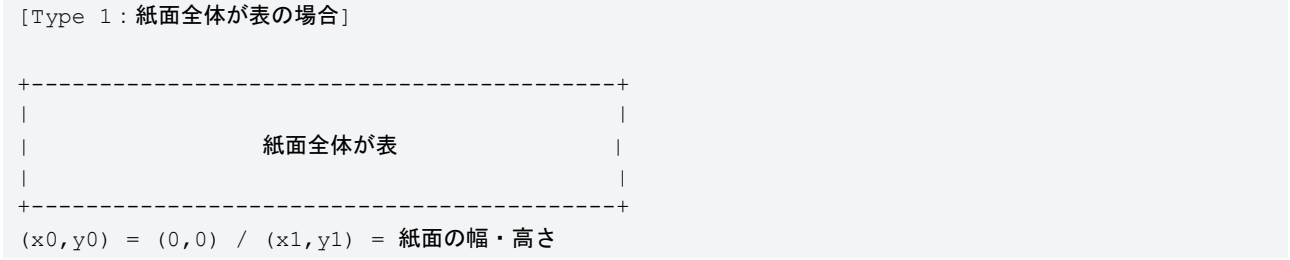
一方で、同じ Minor キーが複数行にまたがる表や、OCR トークンの順序が紙面の見た目と大きく異なる帳票では、Simple の出力だけでは正しい対応関係を判断できない場合があります。その場合は、表組（TypeA） / 表組（TypeB） などの構造化パーサーを使用します。

B.9 正規化データを見るときの確認ポイント

整形前テキストや正規化データを見るときは、まず半角コロンで major / minor / value の位置関係が保たれているかを確認します。次に、項目名が value 側に残っていないか、本来値である文字列が key 側に入っていないか、備考やフッタが本文へ混入していないかを確認します。

表組（TypeA） / 表組（TypeB） などの表組パーサーでは、項目名が value 側に残る場合は rescue_key_list、本来値である文字列が key 側に入る場合は not_key_str_list、OCR 文字の揺れが残る場合は translate_dic を確認します。Simple（テキスト）レイアウトでは、帳票ごとの Simple 用キーワードリストへの追加と、マイナスキーワード指定によって調整します。

レイアウト	主な用途	特徴
Simple	文字の並びが素直な帳票	赤線・青線を使わず、項目名キーワードから Minor / Value の組を作成します。Major は原則取得しません。
表組（TypeA）	1 行に 1 組の項目名と値がある表	Major カラム、Minor カラム、Value カラムを表の横方向で分離します。赤線が重要です。
表組（TypeB）	1 行に複数の Minor / Value ペアが並ぶ表	左側に Major 群、右側に複数の Minor / Value ペアが並びます。赤線と青線が重要です。



表は header / table body / footer に分けて扱われます。
[Type 2：紙面の一部だけが表の場合]

```
+-----------------------------------------+
| top                                     |
|     +-------------------------------+   |
|     |              表               |   |
|     +-------------------------------+   |
|                                  bottom |
+-----------------------------------------+
```

table_top_x, table_top_y, table_bottom_x, table_bottom_y で
表の左上・右下の矩形座標を指定します。
[Simple の考え方]

Minor	Value
物件名	The trias sumiyoshi
所在地	江東区扇橋 3-3-5
地目	宅地
価格	380,800,000 円

出力例：
:物件名:The trias sumiyoshi
:所在地:江東区扇橋 3-3-5
:地目:宅地
:価格:380,800,000 円
[表組（TypeA）：1 行に 1 組の Minor / Value]

table_colmns = Major カラム数（Minor カラムは含めない）

major1	major2	minor	value
土地	面積	公簿面積	29.61 ㎡
法規制		建ぺい率	80%
法規制		容積率	600%

```
                          ^
                          table_split_x（赤線）
```

出力形式：major:minor:value
例：土地-面積:公簿面積:29.61 ㎡

[表組（TypeB）：1 行に複数の Minor / Value ペア]

table_colmns = Major カラム数（ペア数ではない）

major1	major2	minor1	value1	minor2	value2

```
+-----------+-----------+-----------+-----------+-----------+-----------+
| 土地      |           | 所在地      | 横浜市...  | 地目        | 宅地       |
| 法規制    |           | 建ぺい率    | 80%       | 容積率      | 600%      |
+-----------+-----------+-----------+-----------+-----------+-----------+
            ^                       ^           ^
            table_split_x           table_middle_split_x（青線）
            （赤線）                Value 部内の追加分割位置
```

出力形式：`major:minor:value` **をペア数分繰り返す**
例：**土地:所在地:横浜市...**
例：**土地:地目:宅地**
例：**法規制:建ぺい率:**`80%`
例：**法規制:容積率:**`600%`

線	対象パラメータ	基準	指定方法
赤線	table_split_x	OCR トークンの中心 x 座標（cx）基準	クリックした x 座標をマイナス付きで指定します。例：クリック位置が 1807 の場合は -1807。
青線	table_middle_split_x	紙面左端を x=0 とする座標基準	クリックした正の x 座標をそのまま指定します。複数ある場合はカンマ区切りで指定します。例：1878,1879。

帳票の状態	推奨レイアウト	理由
項目名と値が素直に並んでいる	Simple	赤線・青線の調整なしで主要項目を取得できる場合があります。
1 行に 1 つの項目名と値がある	表組（TypeA）	Major / Minor / Value の基本構造で処理できます。
1 行に複数の項目名と値のペアがある	表組（TypeB）	複数ペアを繰り返し処理する必要があります。
地図・写真・ロゴが混在する	表組（TypeA） / 表組（TypeB） + 表範囲指定	table_top_x などで表領域を限定します。

`major:minor:value`

要素	意味
major	大項目。項目の分類や周辺情報を示す参考情報です。複数 Major がある場合は、ハイフンで連結されます。
minor	項目名。CSV 出力で特に重要な項目です。
value	項目の値。CSV 出力される内容です。

形式	意味
major:minor:value	Major、Minor、Value がすべて取得できている通常行です。
major::value	Minor が空欄で、Major と Value だけがある行です。
:minor:value	Major が空欄で、Minor と Value だけがある行です。Simple（テキスト）レイアウトではこの形式が多くなります。

:minor:	Major と Value が空欄で、Minor だけがある行です。
::value	Major と Minor が空欄で、Value だけがある行です。
ヘッダ::	ヘッダー行です。CSV 出力対象外です。
フッタ::	フッター行です。CSV 出力対象外です。
備考:備考:value	備考行です。複数行にまたがる本文を備考としてまとめます。

```
？major:minor:value
major:？minor:value
major:minor:？value
```

確認点	見る内容
項目名と値が正しいか	minor と value の対応を確認します。
Major が空欄かどうか	Major は参考情報のため、空欄でも minor/value が正しければ問題ない場合があります。
低信頼マークがあるか	「？」が付いた値は紙面と照合します。
備考が分断されていないか	備考行は本文がまとまっているか確認します。

Appendix C. 帳票 config パラメーター一覧

本付録では、帳票 config で使用する主なパラメータを一覧で説明します。config は、OCR 結果を紙面どおりに分割・結合するための設定値です。先に OCR エンジンの JSON、行分割データ、整形前テキストを確認し、どの段階で崩れているかを見てから調整してください。

C.1 調整の基本

config を調整するときは、まず OCR 結果画面で、OCR エンジンの JSON、行分割データ、整形前テキストを確認し、どの段階で崩れているかを切り分けます。文字自体の揺れは translate_dic、行や列のまとまりは y_threshold / x_threshold、表の分割位置は table_split_x / table_middle_split_x などで調整します。

注意：キーワード系設定の共通ルール

remarks_list、rescue_key_list、not_key_str_list、key_separator_list、Simple 用キーワード、マイナスキーワードなど、キーワード名・項目名を指定する config 項目では、原則として項目名の途中に半角/全角スペースを入れません。紙面上で文字間が空いていても、設定値は空白なしにします。括弧付き項目名から括弧内だけを指定する場合も、指定する項目名部分は空白なしで登録します。

注意：Simple（テキスト）レイアウトでの調整について

rescue_key_list、not_key_str_list、key_separator_list などは、主に表組（TypeA） / 表組（TypeB） などの表組パーサーで使用する調整項目です。

Simple（テキスト）レイアウトは 1 区画の OCR トークン列を、Simple 用キーワードに従って Minor / Value に分解する方式です。そのため、Simple（テキスト）レイアウトでの調整は、帳票ごとの Simple 用キーワードリストへの追加と、マイナスキーワード指定を中心に行います。

本来キーにしたい語句がキーになっていない場合は、Simple 用キーワードリストに追加してください。本来値にしたい語句がキーになっている場合は、マイナスキーワードに指定してください。

システム内部であらかじめ定義されているキーワードの内容は、一般ユーザーは確認・変更できません。通常の帳票設定では、帳票ごとのキーワード追加とマイナスキーワード指定によって調整してください。

以下の表は、表組（TypeA） / 表組（TypeB） などの表組パーサーで調整する場合の目安です。Simple（テキスト）レイアウトの調整は、前述の Simple 用キーワードリストへの追加とマイナスキーワード指定を使用してください。

現象	主に確認する項目
OCR 文字そのものが間違っている	translate_dic
本来 1 行の内容が複数行に分かれる	y_threshold_rate を大きくします。
本来別行の内容が 1 行に巻き込まれる	y_threshold_rate を小さくします。
横の分割位置がおかしい	table_split_x
Value 部の細分割がおかしい	table_middle_split_x
項目名が value 側に入る	rescue_key_list

値が key 側に入る	not_key_str_list
地図・写真・ロゴ文字が混入する	table_top_x / table_top_y / table_bottom_x / table_bottom_y

C.2 パラメーター一覧

パラメータ	例・初期値	用途・調整ポイント
layout	table_layout_表組（TypeB）	OCR レイアウトパーサー種別名
ocr_dpi	220	OCR 画像の解像度[dpi]（0:自動調整 >0:指定値）
ocr_long_side	1800	OCR 画像の長辺サイズ[px]（0:自動調整 >0:指定値）
ocr_rotate	0	OCR 画像の回転[度]（-1:自動調整 >=0:指定値）
memo	config 設定メモ（config 調整履歴や帳票別の注意事項などの記述用です）	
header	-1	ヘッダーの行数（-1:自動判定 0:ヘッダーなし >0:指定値) ページ上部をヘッダ領域として除外する行数です。 -1: OCR トークン位置から自動判定します。 0 : ヘッダなしとして扱います。 >0: 指定した行数をヘッダとして扱います。 ※自動判定がうまくいかない場合に指定してください。
footer	-1	フッターの行数（-1:自動判定 0:フッターなし >0:指定値) ページ下部をフッタ領域として除外する行数です。 -1: OCR トークン位置から自動判定します。 0 : フッタなしとして扱います。 >0: 指定した行数をフッタとして扱います。 ※自動判定がうまくいかない場合に指定してください。
header_line_include_flag	off	ヘッダーに同じ y 座標の行を含める
header_ratio_starts_list	0.20, 0.10	ヘッダ自動調整の比率リスト(CSV 文字列)（例）0.30, 0.20, 0.10
footer_line_include_flag	on	フッターに同じ y 座標の行を含める
footer_ratio_starts_list	0.90,0.80,0.70,0.60, 0.53, 0.45, 0.40	フッタ自動調整の比率リスト(CSV 文字列)（例）0.80, 0.70, 0.60, 0.53, 0.45, 0.40
footer_keywords_list	株式会社ミライアロー,TEL,FAX,〒,東京都	フッタ構成文字列のリスト(CSV 文字列)（例）株式会社,TEL,FAX,〒,東京都
remarks_list	備考	備考文字列の候補リスト(CSV 文字列)（例）備考 登録時は、項目名・キーワード名の途中に半角/全角スペースを入れないでください。
display_none_value_flag	off	値なし（末尾 :）の通常表を表示する（0:しない 1:する） 値が空の通常行を結果に表示するかを指定します。 1: 表示します。通常運用です。紙面に存在する項目行を結果にも残ます。 0: 表示しません。Simple テキストなどで、不要な値なし行を削除する場合に指定します。
remark_rollback_empty_value_lines	on	直前の値なし（末尾 :）の通常行も、備考として吸収する（表専用 0:しない 1:する） 備考開始直前にある「値なし通常行」も備考へ吸収するかを指定します。 0: 吸収しません。 1: 吸収します。 備考欄は 複数行にまたがって、配置されることが多く、 備考キーが中央付近に印字される紙面では、前方数行が bullet(::value) 行や value なし通常行として出力される場合があります。 この設定を有効にすると、それらの行を備考本文へ吸収します。
remark_max_rollback_rows	5	直前の行を備考として吸収(ロールバック)するかの最大行数（表専用 0:しない >0: 指定行数 -1:すべて） 備考開始直前の行を、最大何行まで備考へ吸収するかを指定します。 0: 吸収しません。 >0: 指定した行数まで吸収します。 -1: 直前の吸収可能行をすべて吸収します。 備考欄は 複数行にまたがって、配置されることが多ため、 備考前方の bullet(::value) 行や value なし通常行を備考へ吸収しないと、 通常は紙面の備考内容と一致しません。
simple_minor_keywords_list		Simple テキストの項目名(minor)候補のリスト(CSV

		文字列) (正規表現も指定可) 登録時は、項目名・キーワード名の途中に半角/全角スペースを入れないでください。
table_top_x	-1	table の左上隅の x 座標[px]（表専用 -1:自動設定(0) >=0:指定値）
table_top_y	-1	table の左上隅の y 座標[px]（表専用 -1:自動設定(0) >=0:指定値）
table_bottom_x	-1	table の右下隅の x 座標[px]（表専用 -1:自動設定(紙面幅) <0:紙面幅-指定値 >=0:指定値）
table_bottom_y	-1	table の右下隅の y 座標[px]（表専用 -1:自動設定(紙面高) <0:紙面高-指定値 >=0:指定値）
translate_dic	中国語フォント変換辞書のカスタム定義(JSON) （例） {"word_variant_map": {"洪谷区":"渋谷区","所有推":"所有権",...}}	"word_variant_map": [] ・カスタム辞書にはこの帳票で「OCR により間違って読み取られた言葉 → 正しい言葉」の形で登録してください ・正しい言葉をさらに別の言葉に変えるような登録はしないでください ・登録する言葉は、できるだけ意味がはっきりした単語で登録してください ・短すぎる言葉（2〜3 文字）は、思わぬ影響が出るため基本的に登録しないでください ・置き換えた後に、元の言葉が残らないようにしてください （同じ変換が何度も繰り返されてしまう原因になります）
table_colmns	1	table の Major カラム数 (表専用 0:自動設定 >0:指定値)（デフォルト値:1）
table_split_x	366	table の Major-Minor 部/Value 部の分割位置[px]（表専用 <0:自動推定 0:自動調整 >0:指定値） 大項目名（Major）- 項目名（Minor）部 と 値（Value）部の分割位置（赤線表示）は、 分割位置の罫線の上でマウスを左クリックして x 座標を拾ってください（クリップボードに入ります） 拾った x 座標を、右クリックで -（マイナス）を付けて指定してください（ペーストされます） table の Major-Minor 部/Value 部の分割位置 が 自動設定されます
table_middle_split_x	713,942	table の Value 部の細分割位置の並び[px](CSV 文字列)（表専用 0:自動設定 >0:指定値 1,指定値 2,,,指定値[Major-Minor 部以降のカラム数-1]） 値（Value）部の細分割位置（青線表示）は、 ペアで構成された 項目名（Minor）と 値（Value）の分割位置を示します 分割位置の罫線の上でマウスを左クリックして x 座標を拾ってください（クリップボードに入ります） 拾った x 座標を、,（カンマ）区切りで右クリックして指定してください（ペーストされます）
tabel_major_minor_reverse	off	table(Matrix)のキー順序(Major-Minor)
y_threshold_row_anchor_mode	0	行分割アルゴリズム（-1:自動調整 0:直前 token 基準 1:行先頭 token 基準 2:行内中央値(可変レンジ) 基準 3:局所行間ピッチ追従 基準）
x_threshold	0	x 座標の閾値[px]（0:自動調整 >0:指定値）（デフォルト値:10）
y_threshold	0	y 座標の閾値[px]（-1:自動推定 0:自動調整 >0:指定値）（デフォルト値:27）
y_threshold_rate	0.4	y 座標の中央値の閾値に対する比率（0:自動調整 >0:指定値）（デフォルト値:0.6) ・確認場所 「OCR パーサーの行分割データ」を開き、同じ rows[n] に入るべきか、別の rows[n] に分かれるべきかを確認してください。 ・本来 1 行の内容が複数の rows[n] に分かれている場合 この値を大きくしてください。 例：0.3 → 0.4 → 0.5 ・本来別行の内容が同じ rows[n] に巻き込まれている場合 この値を小さくしてください。
y_threshold_same_row_hori_rate	0	同一行判定（横結合）の y 座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.0）
y_threshold_same_row_vert_rate	1.4	同一行判定（縦結合）の y 座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.7）
y_threshold_blocktext_rate	0	block_text 判定の y 座標の閾値 に対する比率（0:自動調整 >0:指定値）（デフォルト値:1.2）
key_max_length	0	table のキーの有効最大文字数 (表専用 -1:制限なし 0:自動設定 >0:指定値)（デフォルト値:6）
key_x_threshold	0	table のトークン結合判定 (表専用 0:自動調整 -1:文字数指定 >0:px 指定値）（デフォ

		ルト値:-1）
key_concat_threshold	0	table のトークン結合最大文字数（表専用 0:自動調整 指定値(整数:n, 実数:m.n, 実数:k.k)）
key_minor_threshold	0	table の項目名数(block_text 構造の閾値 (表専用 0:自動設定 >0:指定値)（デフォルト値:1）
block_text_value_x_limit_threshold	0	block_text 構造のバリュー探索の最大距離[px]（表専用 0:自動設定 >0:指定値）（デフォルト値:650）
block_minor_outlier_dx_threshold	0	block_text 構造の左方向外れ項目名の判定距離[px]（表専用 0:自動設定 >0:指定値）（デフォルト値:120）
key_search_mode	3	キー検索アルゴリズム（0:自動調整 1:行内探索 2:拡張探索（value のみ行でも外側キーを探す） 3:（value-only 行では外側キーを探さない））
key_search_rate	0	拡張キー検索範囲の y 座標の閾値に対する比率（0:自動調整 >0:指定値）（デフォルト値:3.0）
major_interpolate_max_rows	2	Major キー補間の検索範囲の最大行数（表専用 0:しない >0:前後の行数 -1:すべて）
key_in_ng_char_list		禁止文字を含むキーの許可リストまたは解除する禁止文字(CSV 文字列) ※禁止文字 ,, .. ※:：/／。、￥¥$＄%％■□・★☆△▽▲▼() （）を含むキーを使用する場合に設定する 通常はキーとして使えない記号付き文字列を、 項目名（major または minor）として使いたい場合に指定してください 禁止文字（例えば、(,)）を指定すると、禁止が解除され (,) をキーに含めることができます
key_split_flag	off	OCR トークン内のキーを分離する（0:しない 1:する） OCR トークン内のキー（救済リストに記載のもの）を分離するかを指定します。 1: 分離します。 OCR トークンが他の文字と結合されている場合に、キーを分離するために使用します。 0: 分離しません。通常運用です。
rescue_key_list		キー文字列の救済リスト(CSV 文字列) ※誤って値(value)側に入った文字列をキーとして救済する場合に設定する 出力行(major:minor:value) の value に、本来は項目名（major または minor）である文字列が入ってしまう場合に指定してください 登録時は、項目名・キーワード名の途中に半角/全角スペースを入れないでください。
not_key_str_list	駅,徒歩,相談,ー	キー文字列の不許可リスト(CSV 文字列) ※誤ってキー(major/minor)側に入った文字列を値(value)として扱う場合に設定する 出力行(major:minor:value) の 項目名（major または minor）に、 本来は値(value)である文字列が入ってしまう場合に指定してください 登録時は、項目名・キーワード名の途中に半角/全角スペースを入れないでください。
key_separator_list		キー文字列の分離文字リスト(CSV 文字列) 登録時は、項目名・キーワード名の途中に半角/全角スペースを入れないでください。
prefix_strip_list		キー先頭の除去文字リスト(CSV 文字列)
noise_token_list		OCR データの中で取り除くべきノイズトークンのリスト(CSV 文字列) OCR 結果から完全に除外したいノイズトークンを CSV 形式で指定します。 地図・ロゴ・写真内文字など、帳票項目ではない文字を指定してください。 例：三角積,Alri Drine,CANA
clip_clear_back_image	off	クリッピング範囲外の画像背景をクリアする（0:しない 1:する）

score_threshold	0.9	スコアの閾値[0~1]
key_list_check	off	表の Major-Minor 部/Value 部の分割位置(tabe_split_x)の決定要求フラグ
key_list		表の Major-Minor 部/Value 部の分割位置(tabe_split_x)の決定候補のキー並び(CSV)（例）'ka','kb','kc' キーの決定候補となる項目名（major または minor）の並びを指定してください ※紙面によっては OCR トークンが罫線に沿ってないものがありますそのような紙面では、マウスによるクリックでは正確な位置を指定できません そのような紙面で、この機能をご利用ください
value_list_check	off	表の Value 部の細分割位置(table_middole_split_x[])の決定要求フラグ
value_list		表の Value 部の細分割位置(table_middole_split_x[])の決定候補リストの値並び(CSV)（例）['va0','vb0','vc0'], ['va1','vb1','vc1'], ... ペアで構成された 項目名（Minor）と 値（Value）の決定候補を [Minor, ...], [Value, ...] の形式の ,（カンマ）区切りで指定してください ※紙面によっては OCR トークンが罫線に沿ってないものがあります
table_rect_check	off	表の配置位置(table_top_x,table_top_y,table_bottom_x,table_bottom_y)の決定要求フラグ

Appendix D. OCR 揺れに関する変換辞書について

本付録は、OCR 誤認識・中国語フォント類似文字・文字化けに対して、translate_dic に登録する変換辞書の考え方と登録ルールを整理した技術者向け資料です。帳票 config を作成・調整する担当者は、OCR 結果を確認したうえで、辞書で直すべき内容と、パーサー処理や config 調整で直すべき内容を切り分けてください。

D.1 translate_dic（帳票別カスタム辞書）の位置づけ

帳票 config の translate_dic は、この帳票でのみ有効な OCR 揺れ補正用のカスタム辞書です。ユーザーが画面から登録・編集できるのは word_variant_map だけです。

word_variant_map には、OCR により間違って読み取られた語句をキーにし、正しい語句を値として登録します。登録の基本は「OCR 誤認識 → 正規形」の一方向変換です。

システム全体に適用される辞書や文字単位の補正はシステム側で管理されるため、帳票 config のカスタム辞書では扱いません。

```
{
  "word_variant_map": {
    "?口月額賃料": "?月額賃料",
    "惠比券": "恵比寿駅",
    "格松額": "価格総額"
  }
}
```

D.2 word_variant_map の基本登録ルール

word_variant_map は、プログラム側でキー長の長い順に自動ソートしてから適用します。そのため、辞書登録時に並び順を意識する必要はありません。ただし、変換は文字列が安定するまで反復適用されるため、辞書キーと置換後文字列の関係には注意が必要です。

- 辞書は OCR 誤認識 → 正規形 の一方向だけを登録します。
- 正規形 → 別の正規形、のような連鎖前提の登録は避けます。
- 帳票 config のカスタム辞書には、語として補正する必要があるものを word_variant_map に登録します。

D.3 登録してよい語

登録してよいのは、OCR 誤認識語がそのまま正規形に置換され、置換後文字列に置換前キーがそのまま残らないものです。

```
OK 例:
  "惠比券": "恵比寿駅"
```

"格松額": "価格総額"
"東急束横棕": "東急東横線"

D.4 登録してはいけない語

反復変換で自己増殖する定義は禁止します。置換後の文字列に元キーが含まれると、反復適用により同じ文字が増殖するためです。

NG 例:
"東京メトロ日比谷": "東京メトロ日比谷線"

この定義では、置換後の「東京メトロ日比谷線」の中に、元キー「東京メトロ日比谷」が再び含まれます。
その結果、反復適用により次のように増殖します。

東京メトロ日比谷
→ 東京メトロ日比谷線
→ 東京メトロ日比谷線線
→ 東京メトロ日比谷線線線

意味のない短い文字列も、何回目かの反復で自己合食や誤置換を起こす可能性があるため避けます。

NG 例:
"目地" → "地目"

短い語（2〜3 文字）の登録は原則禁止です。ただし、単独で意味が確定し、他語に含まれず、反復変換で問題がない場合に限り、汎用化のために登録できます。

D.5 反復変換で安全な定義

反復変換で安全にするためには、置換後に元キーが残らない形にします。必要に応じて、より具体的な長い語で登録します。

安全な登録例:
"東京メトロ日比谷『": "東京メトロ日比谷線『"
"【恵比寿駅]駅": "『恵比寿』駅"

D.6 長文辞書を残す条件

短い語へ分解しても、現在の反復変換ロジックで 1 件に収束しないものは、長文のまま登録して構いません。長文辞書は汎用性は下がりますが、帳票内の特定の誤認識を安全に補正できる場合があります。

長文辞書の例:
"JR 山手線･東京メトロ日比谷『恵比寿』駅"
→ "JR 山手線・東京メトロ日比谷線『恵比寿』駅"

D.7 実務上の判断基準

新しい語を追加するときは、次の順で判断します。

・word_variant_map の短い語で直せるかを確認します。

・反復変換で正しく収束するかを確認します。

・収束しない場合のみ、長文辞書を残します。

D.8 構造付き辞書の登録ルール

word_variant_map では、通常の単語だけでなく、正規化済み行データの構造を利用した辞書登録を許可します。構造付き辞書とは、":" を含む行データ上の major / minor / value 境界を利用して、通常単語よりも安全に誤認識を補正する辞書です。

```
構造付き辞書の例:
  "利回り:?面利回り:" → "利回り:?表面利回り:"
  "利回り:面利回り:"  → "利回り:表面利回り:"
```

登録してよい条件は次のとおりです。

・":" を含み、通常文章ではなく OCR 正規化行の構造に限定されること。

・置換対象が minor / key / value の境界を含み、他語と誤マッチしにくいこと。

・置換後に元キーが残らず、反復変換で自己増殖しないこと。

・OCR 誤認識 → 正規形 の一方向であること。

・短い語を補正する場合でも、":" により構造限定できること。

登録してはいけない条件は次のとおりです。

・置換後に置換前キーが残るもの。

・":" を付けても、他の正常行に広くマッチするもの。

・value の内容まで過度に巻き込んで、紙面値を書き換えるもの。

・出力構造そのものを別構造に作り替える目的のもの。

構造付き辞書は、ロジックで帳票依存補正を追加するより安全な場合に使います。ただし、本来 output_lines() 側で削除・結合・補間すべき行を、辞書だけで無理に解決してはいけません。

D.9 構造付き辞書の例と注意点

```
OK 例:
  "現況賃料（月:" → "現況賃料（月額）:"
  "現況賃料（月:?管理月:" → "現況賃料（月額）:?管理費（月額）:"
  "収支現況賃:?(部):" → "収支:?現況賃料（年額）:"
  "収支現況賃:?管理費等:" → "収支:?管理費等（年額）:"

要注意例:
  "収支現況賃:?(年）:" → "収支:?現況賃料（年額）:"
```

要注意例は、項目名補正としては可能ですが、value なし行です。最終的には erase_none_value_lines() 側で削除されるべき行であり、辞書だけで無理に帳票構造を成立させるべきではありません。

D.10 辞書追加後の確認

辞書を追加した場合は、必ず実際の帳票で OCR 再実行を行い、正規化データとキーワード解析結果を確認します。特に、短い語・構造付き辞書・長文辞書を追加した場合は、別帳票への副作用がないかも確認してください。

Appendix E. CSV 出力時の項目展開と特殊項目について

本付録では、キーワード解析結果画面で選択された Excel 項目が、Excel / CSV 出力時にどのように扱われるかを説明します。

E.1 Excel 項目マスターと出力項目

キーワード解析結果画面では、OCR で抽出された minor / value に対して、出力先となる Excel 項目が割り当てられます。Excel 項目には、通常の 1 対 1 対応の項目のほか、出力時に分解・展開される特殊項目があります。画面上では 1 つの値として表示されていても、Excel / CSV 出力時には複数の実項目へ分解される場合があります。

E.2 出力時の扱いの分類

分類	扱い
通常項目	OCR 値を、対応する Excel 項目へ 1 対 1 で出力します。
未設定	Excel / CSV には出力しません。不要な行は未設定にすることで出力対象から外せます。
重複項目	同じ出力項目へ複数値が割り当てられた場合、スペース区切りで結合して出力します。
仮受け項目	交通、バス、接道、建ぺい率／容積率など、複数要素を一時的に受け、出力時に展開します。
分解項目	住所や年月など、出力時に複数の実項目へ分解される項目です。

E.3 未設定の扱い

自動マッピングで出力先を判断できない場合、または利用者が意図的に出力対象から外したい場合は、対応する Excel 項目を「未設定」にします。未設定の内容は Excel / CSV には出力されません。

E.4 重複項目の扱い

同じ Excel 出力項目に複数の OCR 値が割り当てられた場合は、重複として扱われます。重複した内容は捨てられません。Excel / CSV 出力時には、同じ項目の値としてスペース区切りで結合されます。

E.5 丸数字付き仮受け項目

丸数字付きの項目は、OCR から取得した複数要素の情報を一時的に受けるための仮受け項目です。主に次の系統で使用されます。

系統	例	説明
交通	交通①　交通②　交通③	路線名、駅名、徒歩分数などをまとめて受け、出力時に必要項目へ展開します。

バス	バス①、バス②、バス③	バス路線・停留所・徒歩分数などを受けます。
接道	接道①、接道②、接道③	方位、種別、幅員などの道路情報を受けます。
建ぺい率／容積率	建 4①、建 4②	建ぺい率と容積率の組を受けます。この系統は「建 4」の 2 文字と丸数字の形式で表示されます。

同じ種類の情報が複数取得された場合は、原則として番号が順番に上がります。用意されている番号数を超える情報や、同じ出力項目に割り当てられた情報は、重複として結合出力されます。

E.6 住所系の分解

住居表示などの住所情報に都道府県名が含まれている場合、Excel / CSV 出力時に都道府県名を別フィールドへ分離する場合があります。この場合、住所本体の項目には都道府県名を除いた住所が入り、都道府県用の項目には分離された都道府県名が入ります。

E.7 年月系の分解

築年月、竣工年月などの年月情報は、画面上では 1 つの値として表示される場合があります。Excel / CSV 出力時には、必要に応じて年と月へ分解され、基幹システム側の年項目・月項目へ出力されます。

例：平成 3 年 1 月 → 年項目：平成 3 年 / 月項目：1 月

E.8 インポート対象とインポート外項目

画面下部の「EXCEL 出力対象」では、出力する項目の範囲を選択できます。

出力対象	説明
すべての項目	インポート対象とインポート外項目の両方を出力します。
インポート項目のみ	基幹システムへ取り込む実項目のみを出力します。
インポート外項目のみ	仮受けや確認用など、基幹システムへ直接取り込まない項目のみを出力します。

E.9 出力前の確認ポイント

確認点	内容
未設定が残っていないか	本来出力すべき内容が未設定になっていないか確認します。
重複が妥当か	重複内容が結合されても問題ないか確認します。
仮受け項目が正しいか	交通、バス、接道、建 4 系の割り当てを確認します。
Value に誤読がないか	ペンアイコンで修正できる値は出力前に修正します。

改訂履歴

Ver	日付	内容
1.0.0	2026 年 6 月 27 日	初版作成。