DH ・ 技術要素 ・ ALTO / PAGE

ALTO・PAGE入門

OCRの結果を「座標つき」で残す

DH入門 ／ 技術要素シリーズ

中村

※実験的な取り組みです（構成・図・AI音声合成を含む）。内容はご確認・ご注意のうえご利用ください

この動画について

- ✓ **オープンに公開された仕様・資料**を参照し、独自に構成した解説です
- ✓ スライド・図は新規作成、ナレーションは**AI音声合成**（この回は本人のクローン声ではありません）
- ✓ これは**実験的な取り組み**です。内容は**ご確認・ご注意のうえ**ご利用ください
- ✓ 誤りに気づいたら概要欄からご指摘ください。出典・ライセンスは末尾と概要欄に記載しています

この回のゴール

OCRの結果を、ただの文字列でなく「画像のどこに何があったか」ごと残す考え方をつかむ

- ✓ OCR・HTRの結果を**座標つき**で残す意味を説明できる
- ✓ **ページ→領域→行→単語**という入れ子の構造をイメージできる
- ✓ **ALTO** と **PAGE** が、それぞれどんな場面で使われるか見当がつく
- ✓ 座標つきだから**画像とテキストを結びつけて使える**（検索・修正・TEI/IIIFへ橋渡し）と説明できる

XMLやTEI・IIIFを知っていると分かりやすいですが、必須ではありません。

今日の流れ

✓ **なぜ「座標つき」で残すのか** — ただの文字列では捨ててしまうもの

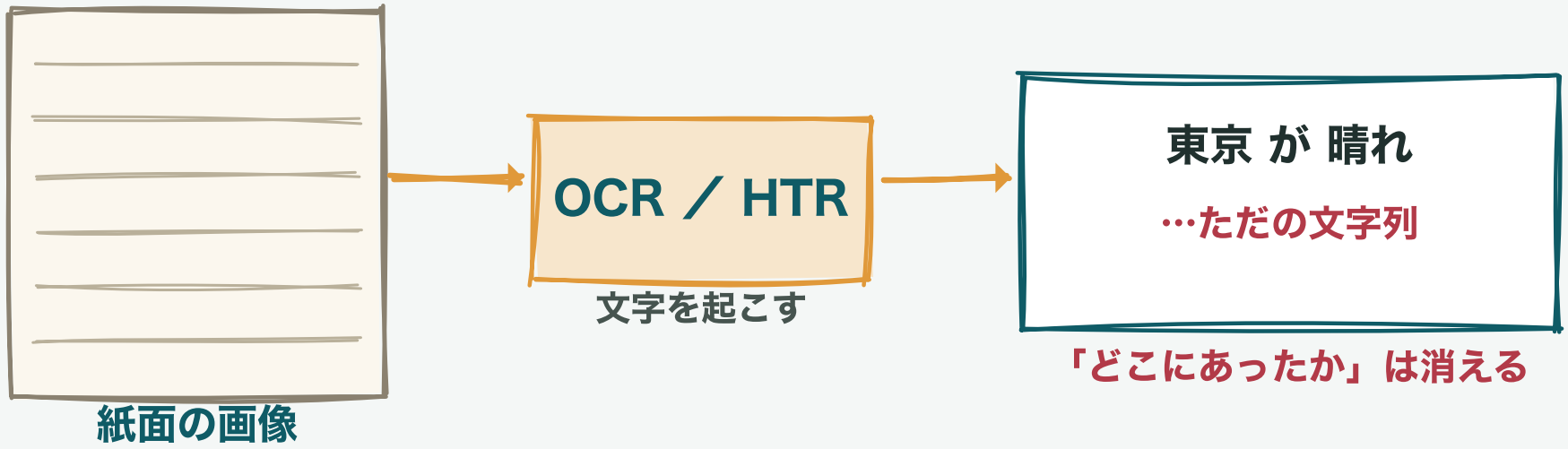
✓ **入れ子の構造** — ページ・領域・行・単語、そして座標

✓ **ALTO と PAGE** — 二つの標準の性格の違いと、その先の活用

1. なぜ「座標つき」で残すのか

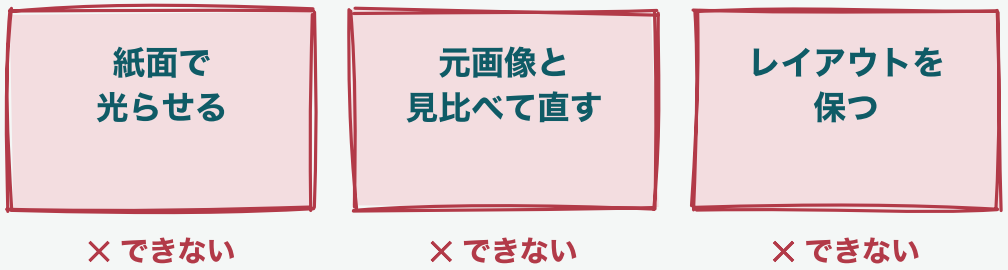
ただの文字列にすると、何が捨てられてしまうのでしょうか

画像から文字を起こす — その結果は？



OCR（活字）や **HTR**（手書き）は、画像から文字を起こします。でも「ただの文字列」にすると、**紙面のどこにあったか**は消えてしまいます

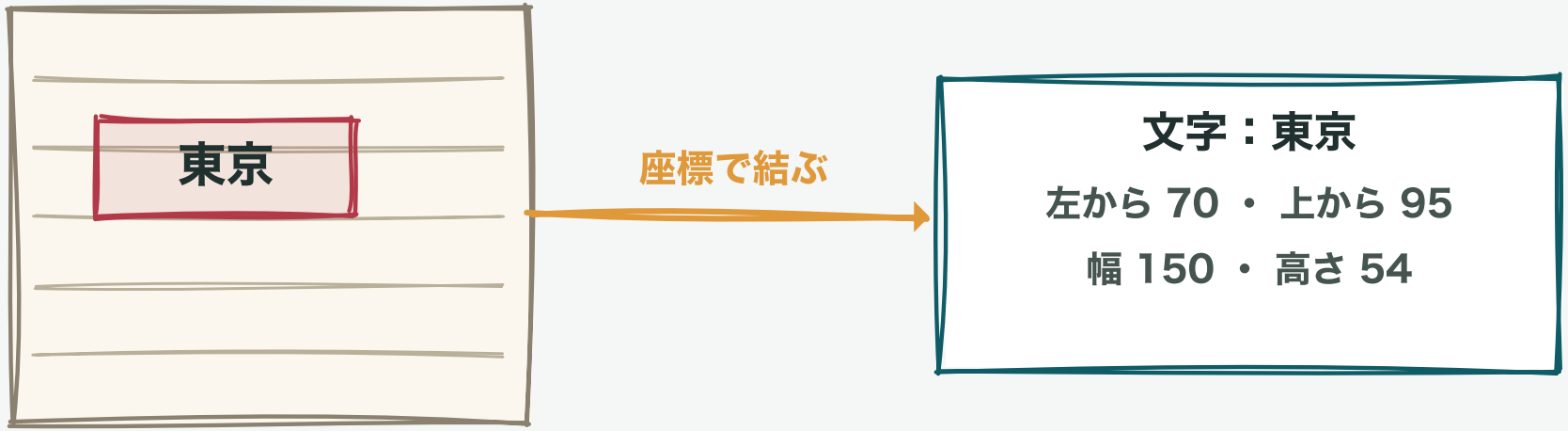
位置を捨てると、できなくなること



- ✓ 検索した語を**紙面の上で光らせる**
- ✓ 誤認識を**元画像と見比べて直す**
- ✓ 段組み・見出しなど**レイアウトの構造**を保つ

これらは、文字が**どこにあったか**を一緒に持っていないと、難しくなります

座標が、画像とテキストを結ぶ



そこで、文字ごとに**座標**（左から・上から・幅・高さ）を添えます。すると、画像のこの位置に、この文字、と**結びつけて**残せます

ここまでの整理

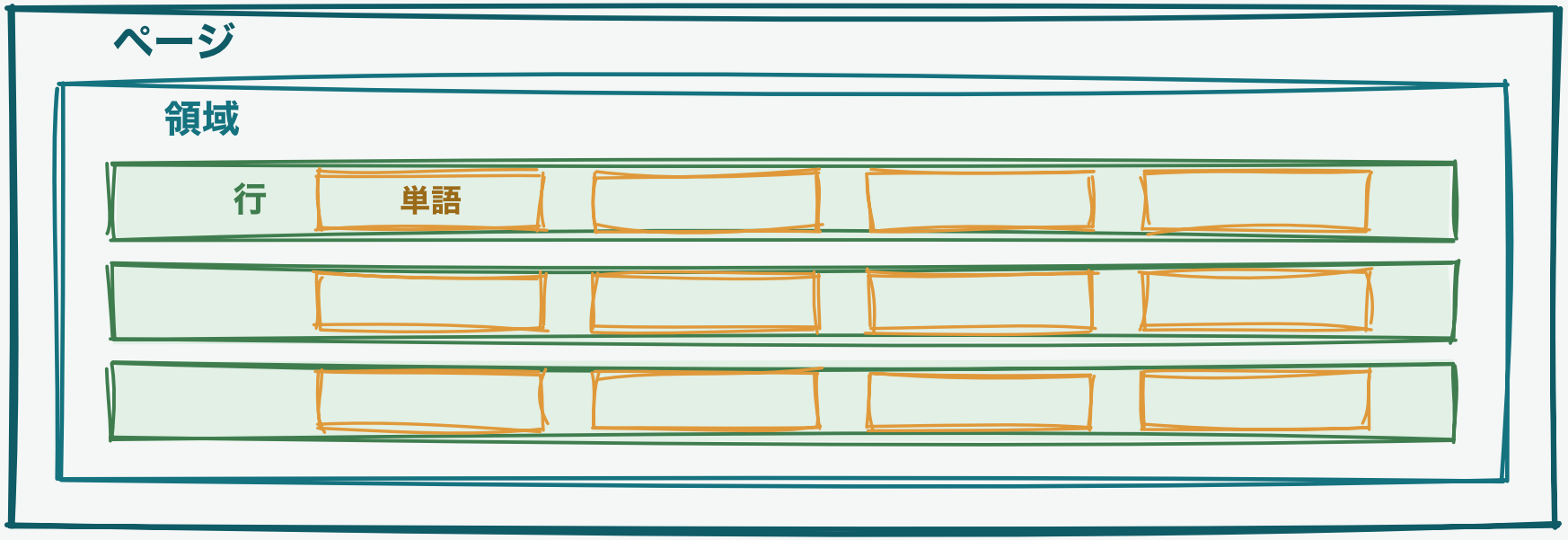
- ✓ OCR・HTRの結果を「ただの文字列」にすると、**位置の情報**が失われる
- ✓ 位置がないと、**検索のハイライト・修正・レイアウト保持**が難しくなる
- ✓ 文字に**座標**を添えると、画像とテキストを**結びつけて**残せる

では、その「座標つきの結果」を、どんな形で書き表すのでしょうか

2. 入れ子の構造

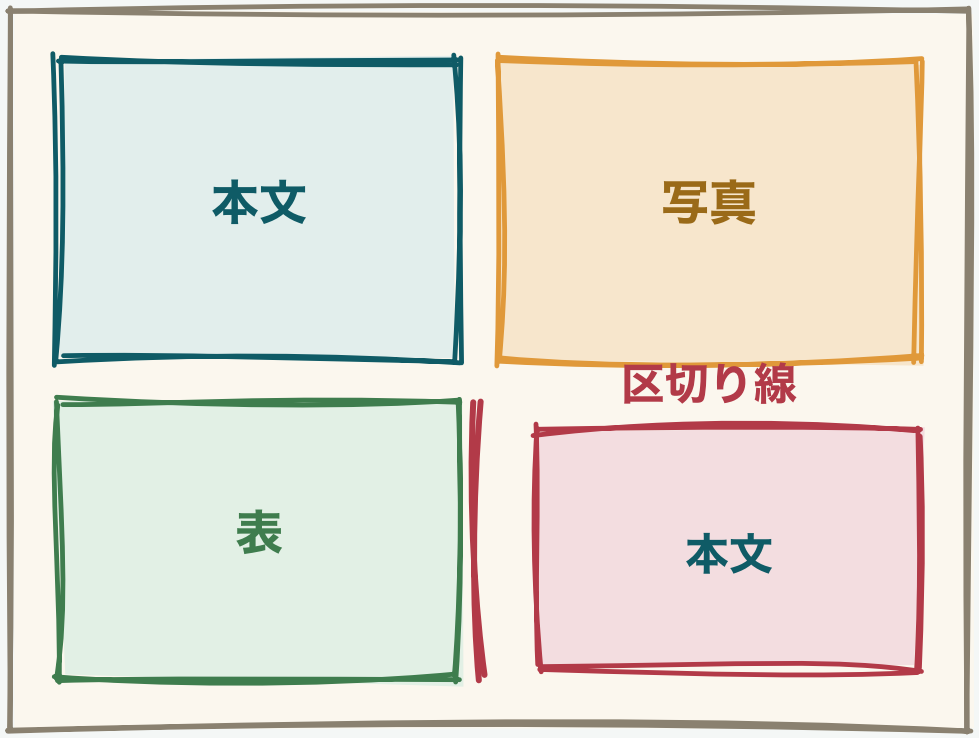
ページ・領域・行・単語 ― 紙面を、入れ子で写し取ります

紙面を、入れ子でとらえる



紙面は、**ページ**の中に**領域**、領域の中に**行**、行の中に**単語**、という**入れ子**でとらえられます。各段に座標がつきます

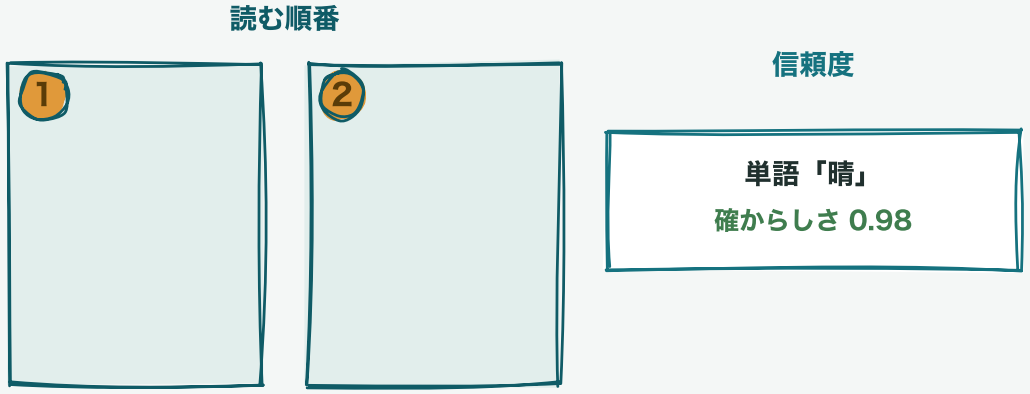
領域には「種類」がある



- ✓ **本文**のかたまり（テキスト領域）
- ✓ **写真・図版**の領域
- ✓ **表**や**区切り線**の領域

本文だけでなく**写真・表・区切り**も区別する。これが**レイアウト解析**です

「読む順番」と「自信のほど」も残せる



✓ **読む順番**（多段組みをどうたどるか）

✓ 認識の**確からしさ（信頼度）**

座標に加えて、**読む順番**や、文字ごとの**信頼度**も残せます。信頼度が低い所だけ、人が見直す、といった使い方ができます

ここまでの整理

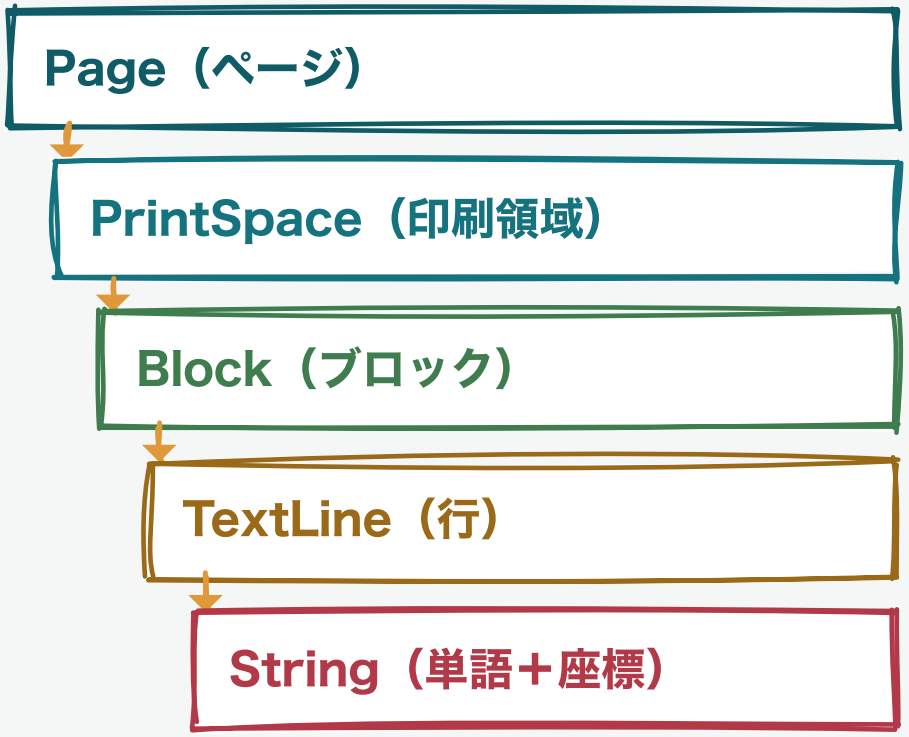
- ✓ 紙面は**ページ→領域→行→単語**の入れ子でとらえ、各段に座標がつく
- ✓ 領域には**種類**（本文・写真・表・区切り）があり、これを**レイアウト解析**という
- ✓ **読む順番**や**信頼度**も一緒に残せる

この入れ子を、実際に書き表す標準が、二つあります。ALTO と PAGE です

3. ALTO と PAGE

同じことを書く、二つの標準。その性格の違いを見ます

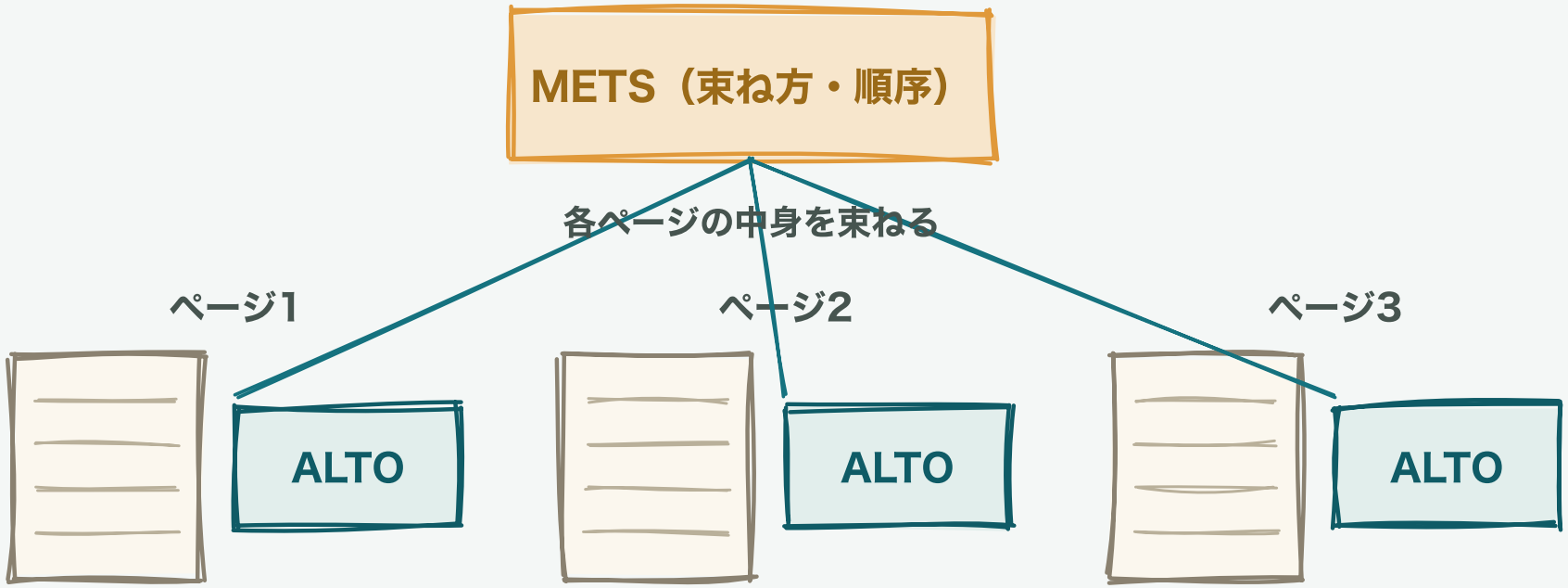
ALTO — 電子化の現場から



- ✓ 図書館の**電子化**の現場で広く普及
- ✓ ページ→印刷領域→ブロック→行→**単語**の階層
- ✓ 各**単語**に座標と認識文字、信頼度

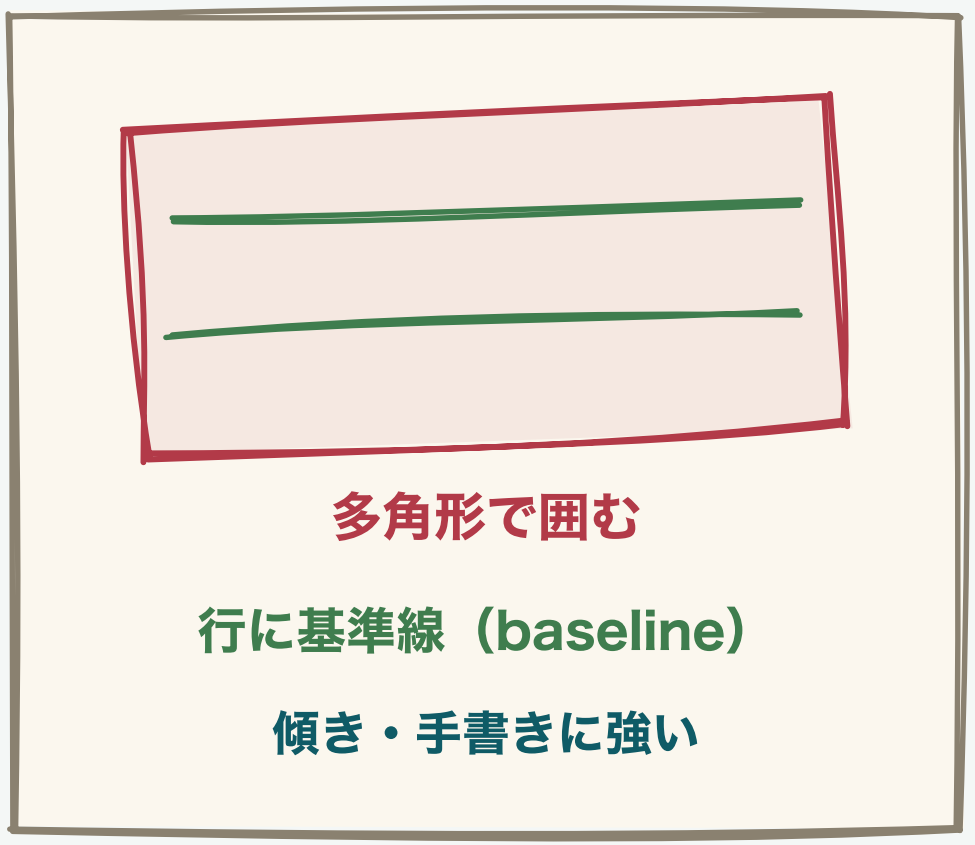
ALTO（アルト）。図書館・新聞の電子化で定着した形式で、現在は**米国議会図書館**が仕様を維持しています

ALTO は「中身」、METS は「束ね方」



電子化では、**ALTO** が各ページの中身（文字と座標）を、**METS** が全体の束ね方（順序・構造）を受け持つ、という組み合わせが定番です

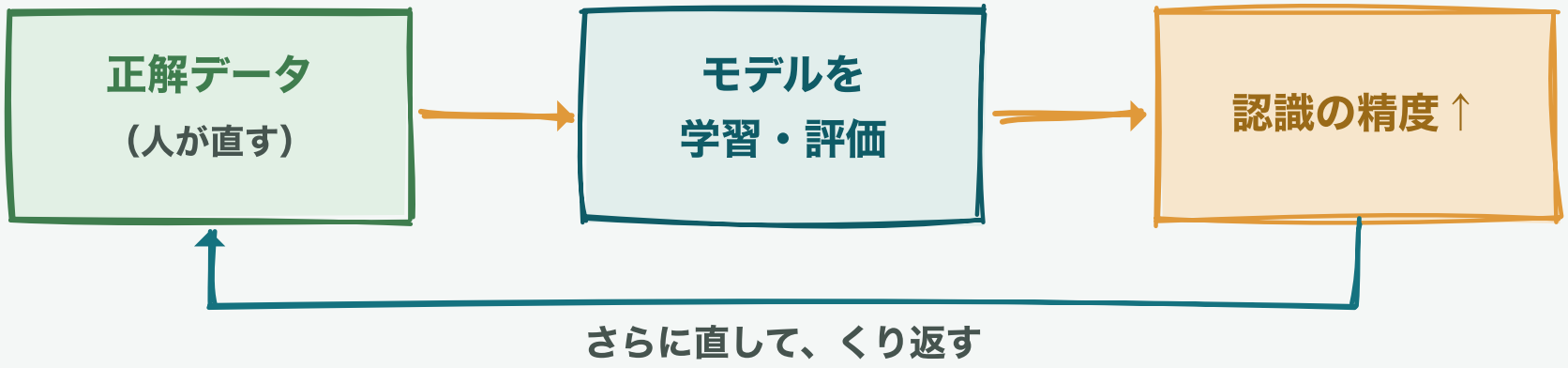
PAGE — 研究・正解データから



- ✓ レイアウト解析の**研究**から生まれた形式
- ✓ 領域を**多角形**で囲める（傾き・曲がりに強い）
- ✓ 行の**基準線**など、手書き資料に向く

PAGE（ページ）。研究機関 **PRImA** 由来で、**正解データ**づくりや手書き認識（Transkribus 等）でよく使われます

「正解データ」を作って、認識を鍛える



PAGEは、人が手で正しく直した**正解データ**を残すのに向きます。これでモデルを**学習・評価**すると、認識の精度を高めていけます

二つの性格 — どちらも「座標つきの結果」

ALTO

電子化・公開で普及

METS と組み合わせる

四角い領域が中心

議会図書館が維持

PAGE

解析・正解データで普及

多角形・基準線が使える

手書き資料に強い

PRImA 研究室が由来

共通：ページ→領域→行 の「座標つき入れ子」

役割は重なりますが、**ALTO は電子化と公開**、**PAGE は解析と正解データ**に強い、と捉えると分かりやすいです。**道具が相互に変換**することもあります

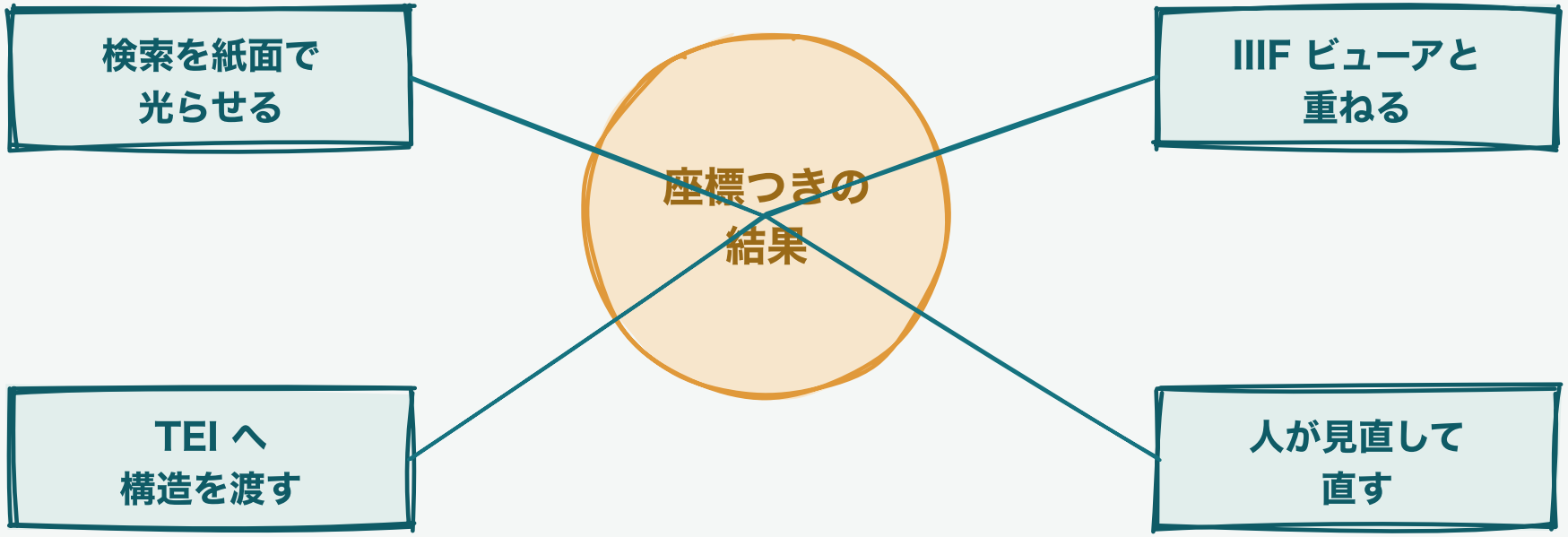
ここまでの整理

- ✓ **ALTO**：図書館の電子化で普及。**METS** と組んで公開に使われる
- ✓ **PAGE**：研究由来。**多角形・基準線**で手書きや**正解データ**に向く
- ✓ どちらも**座標つきの入れ子**。目的に応じて選び、変換もできる

最後に、座標つきだからこそ開ける「使い道」を見ておきましょう

その先の使い道

座標つきだから、つながる



座標つきで残しておくと、**検索結果を紙面で光らせる**、**IIIF** の画像ビューアと重ねる、**TEI** へ構造を引き継ぐ、といった先へ**つなげて**いけます

考えてみよう

ここで少し、考えてみよう

いちど動画を止めて、考えてみてください。

- ✓ あなたが扱いたい資料（新聞・古典籍・手書き史料…）は、**四角い領域**でうまく囲めそうですか。それとも**多角形**が要りそうですか
- ✓ その資料では、**読む順番**はどれくらい複雑でしょうか

資料の性格が、ALTO と PAGE のどちらに向くかの**ヒント**になります

まとめ

- ✓ OCR・HTRの結果を**座標つき**で残すと、画像とテキストを**結びつけて**使える
- ✓ 紙面は**ページ→領域→行→単語**の入れ子。種類・読む順番・信頼度も残せる
- ✓ **ALTO**＝電子化・公開（METSと組む）／**PAGE**＝解析・正解データ（多角形・基準線）
- ✓ 座標つきだから、**検索・IIIF・TEI** へとつなげていける

OCRの結果は「ゴール」ではなく、活用への**出発点**。座標が、その橋渡しをします

出典・ライセンス

本動画の**スライド・図・ナレーション原稿は CC BY 4.0** で公開します（© 2026 中村 覚）。出典表示のうえ自由に再利用いただけます。

- ✓ 参照（事実確認・翻案せず）：ALTO 公式仕様 ／ 米国議会図書館（loc.gov/standards/alto）
- ✓ 参照（事実確認・翻案せず）：PAGE XML 仕様 ／ PRImA Research Lab（primaresearch.org）

図はいずれも概念のみを参照した**新規作画**です。掛け合い版の音声・立ち絵は VOICEVOX／坂本アヒル氏の各規約に従います。

ご清聴ありがとうございました