DH ・ 技術要素 ・ Webアーカイブ

Webアーカイブ入門

消えるWebを、WARC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

消えていくWebを「丸ごと」残す、という考え方をつかむ

- ✓ Webページが**消える・変わる**ため、丸ごと残す必要があると説明できる
- ✓ Webアーカイブの**三段階**（集める・残す・再生する）をイメージできる
- ✓ **WARC** が、通信のやりとりごとWebを束ねて残す入れ物だと説明できる
- ✓ Webアーカイブが**長期保存・改変検知**とつながっていると見当がつく

長期保存（OAIS）や fixity 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。

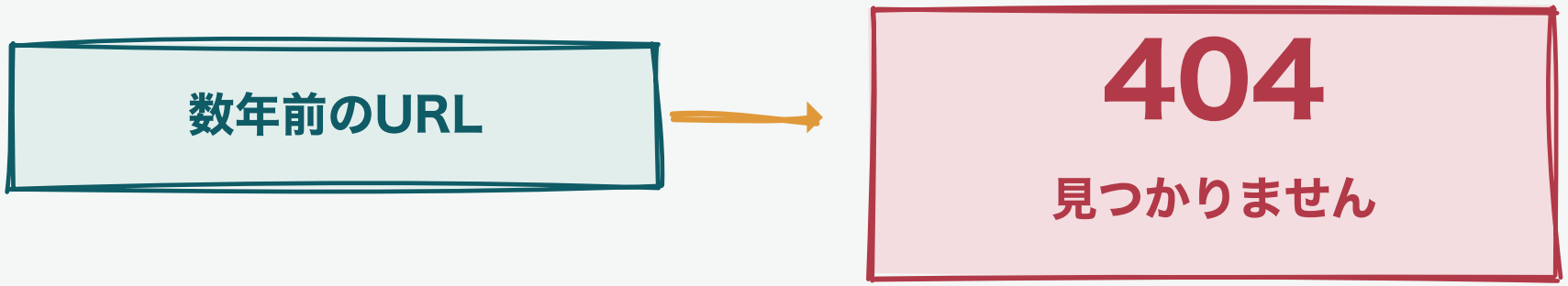
今日の流れ

- ✓ **なぜ消えるのか** — リンク切れと、書き換わる内容
- ✓ **三段階とWARC** — 集める・残す・再生する、その「残す」器
- ✓ **WACZと長期保存** — 持ち運ぶ新しい器、そして残し続ける仕組み

1. なぜ消えるのか

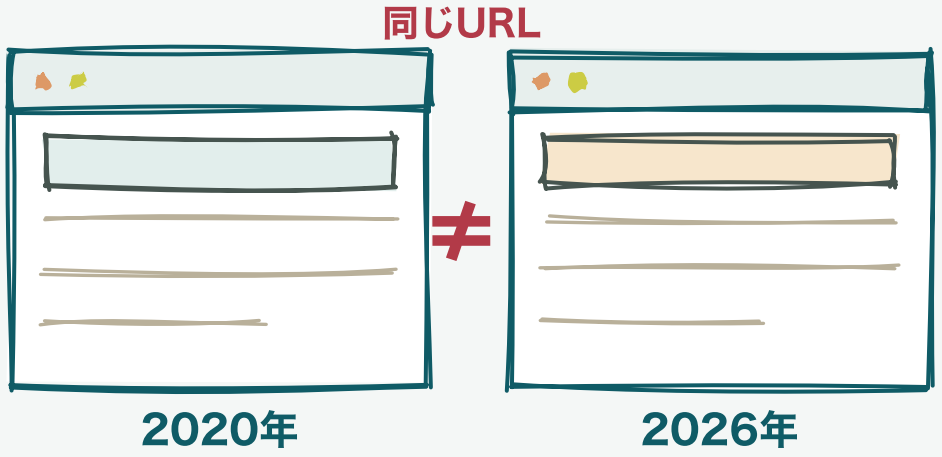
Webは、いつでもそこにある ― とは限りません

クリックしたら、もう無い



数年前のページや、論文に引用したリンクの先が、**消えている**。これは**リンク切れ**と呼ばれ、Webではごくふつうに起こります

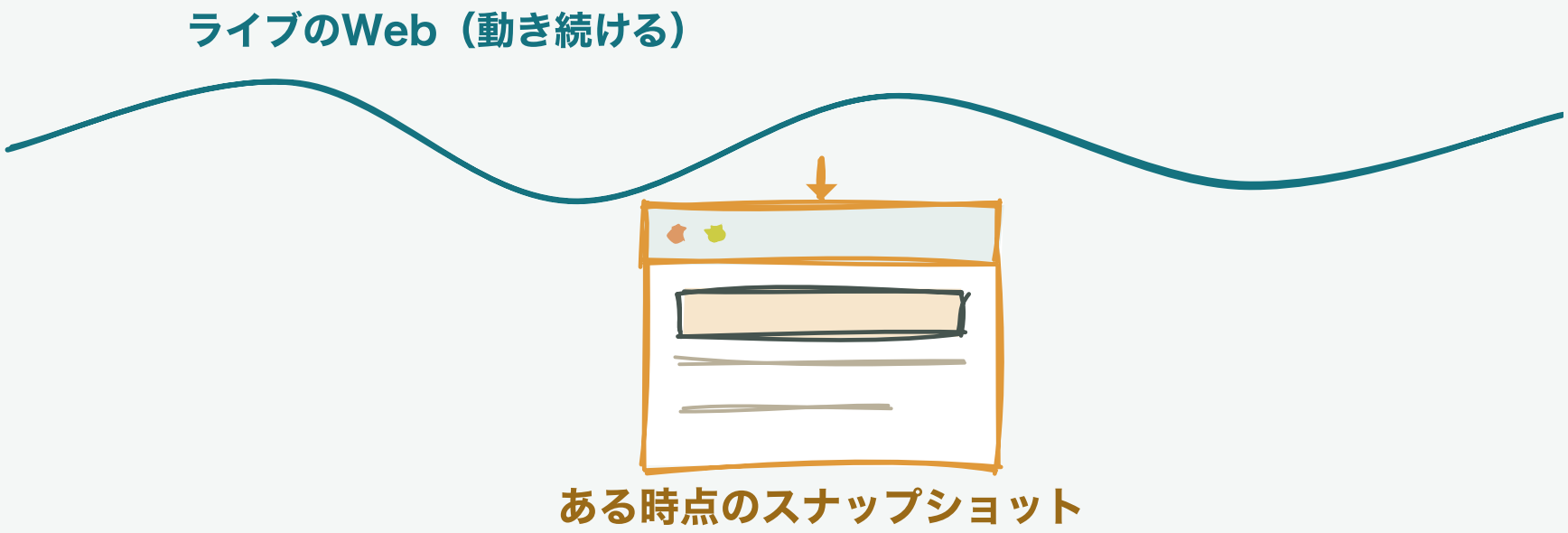
消えなくても、中身が変わる



- ✓ URLは**同じ**なのに、中身が違う
- ✓ 更新・削除・差し替えは**痕跡が残らない**
- ✓ 「あの時こう書いてあった」を**示せない**

たとえページが残っても、内容は**書き換わって**いきます。これを**内容の移ろい**といいます

だから「その時点」を、丸ごと残す



Webは流れ続ける動く対象です。研究や記録のためには、**ある時点のすがた**を、まるごと切り取って残しておく必要があります

ここまでの整理

- ✓ Webは、リンクの先が**消える**（リンク切れ）
- ✓ 残っていても、同じURLの**中身が書き換わる**（内容の移ろい）
- ✓ だから、**ある時点**のすがたを丸ごと残す必要がある

では、その「丸ごと残す」は、どんな手順と器でおこなうのでしょうか

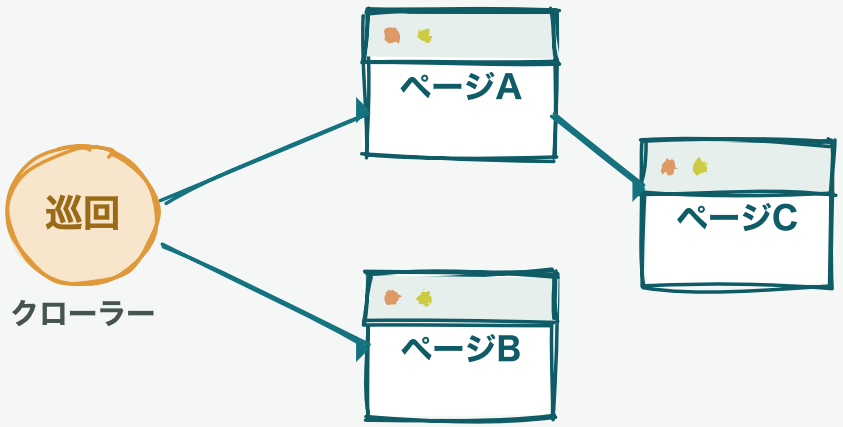
2. 三段階と WARC

集める・残す・再生する ― その「残す」器が WARC です

Webアーカイブの三段階

Webアーカイブは、大きく三段階です。**集める（クロール）**、**残す（WARC）**、そして後から**再生する**。今日の主役は、まん中の「残す」器です

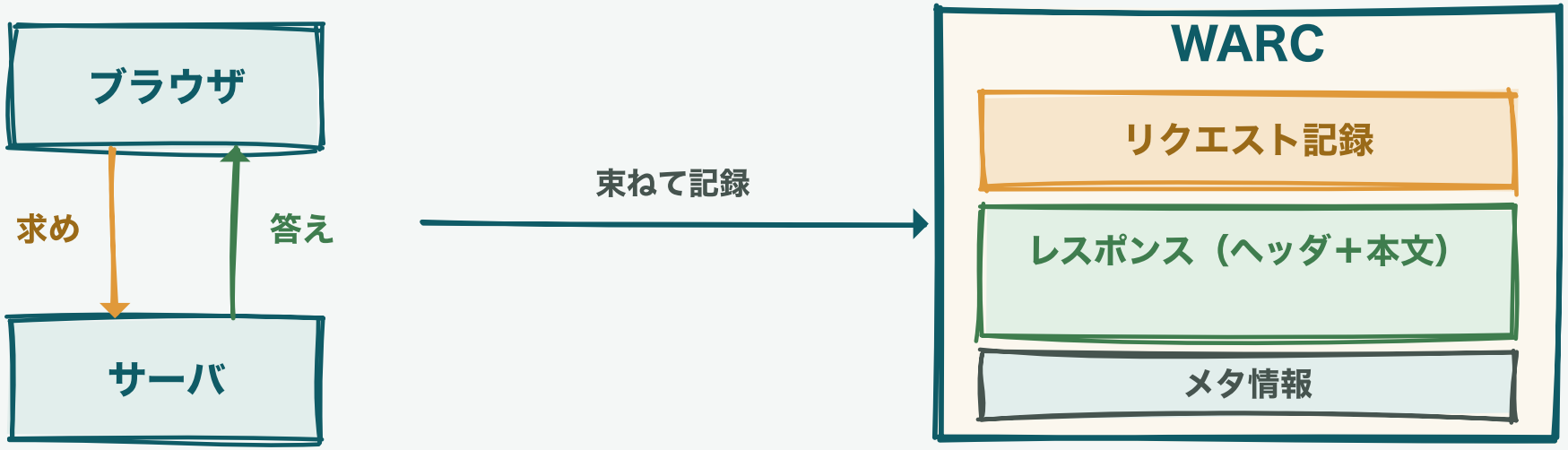
集める — クローラーがたどる



- ✓ 起点から**リンクをたどって**巡回する
- ✓ HTMLだけでなく**画像・CSS**も集める
- ✓ 道具の例：Heritrix、Browsertrix

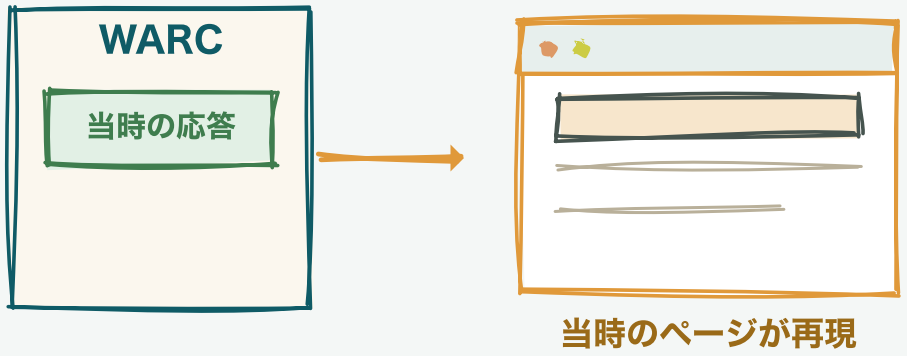
まず**クローラー**という道具が、リンクをたどってページを巡り、表示に必要なものを**ひと通り**集めます

WARC — 通信のやりとりごと束ねる



スクリーンショットではありません。**WARC** は、ブラウザの**求め（リクエスト）**とサーバの**答え（レスポンス）**を、生のまま記録として束ねます

なぜ「やりとり」ごと残すのか



- ✓ 後から**同じ応答**をブラウザに返せる
- ✓ 当時のページを**そのまま再生**できる
- ✓ 見た目だけでなく**中身も**そろう

応答そのものを残すので、後でブラウザに返せば、**その時のページ**を再現できます。画像だけの保存とは、ここが違います

再生する — 当時のWebに戻る

残した WARC を**再生ソフト**に渡すと、**当時の日付**のページがブラウザによみがえります。
Wayback Machine などが、この役目を担います

ここまでの整理

- ✓ Webアーカイブは**集める・残す・再生する**の三段階
- ✓ **クローラー**がリンクをたどって、表示に必要なものを集める
- ✓ **WARC** は、求めと答えのやりとりごと束ねて残す（応答そのもの）
- ✓ だから後から**その時のページを再生**できる

この WARC を、もっと持ち運びやすくした新しい器と、長く残す仕組みを見ます

3. WACZ と 長期保存

持ち運ぶ新しい器と、残し続ける仕組みへ

WACZ — 持ち運びやすい器

✓ WARC に**索引・ページ一覧**を添える

✓ まとめて**ひとつのzip**に

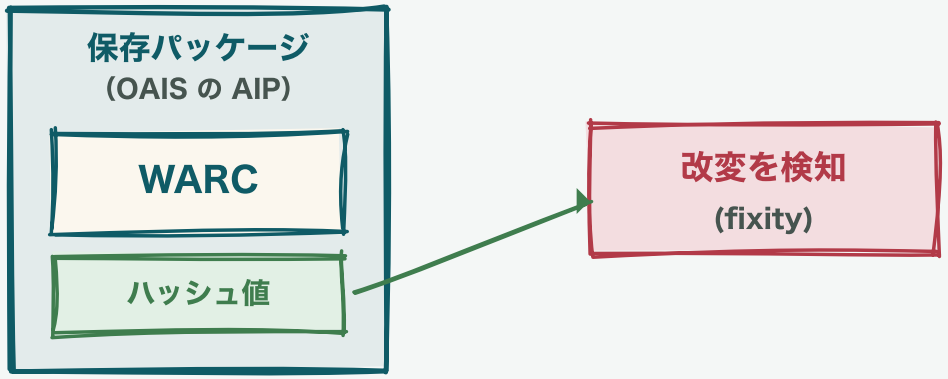
✓ その場で**開いて再生**しやすい

WACZ（ワックゼット）は、WARC に索引などを添えて zip でまとめた、より**持ち運びやすい**新しい器です

みんなで支える標準

WARC は **国際規格（ISO）** として定められ、**IIPC** という国際組織が維持しています。世界の図書館やアーカイブ機関が、協力して支えています

残した後も、守り続ける



✓ WARC は長期保存の**中身**になりうる

✓ **ハッシュ**で改変を検知（fixity）

✓ OAIS の考え方で**残し続ける**

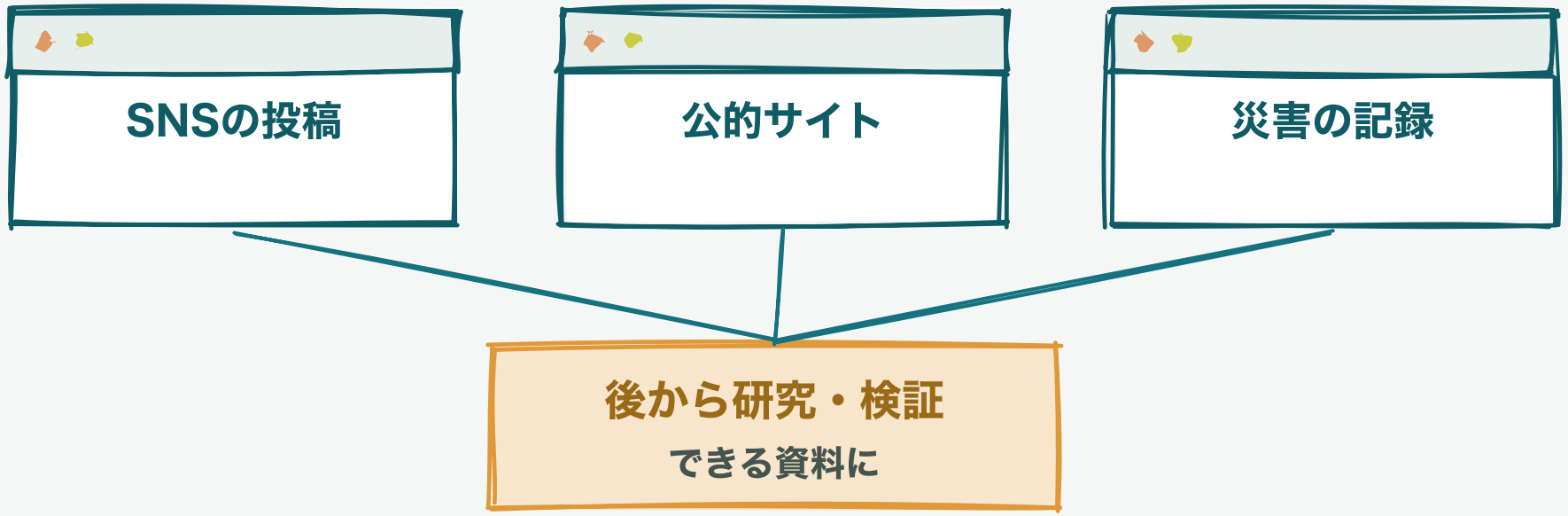
残して終わりではありません。**長期保存（OAIS）**の枠組みに収め、**改変検知（fixity）**で守り続けます

ここまでの整理

- ✓ **WACZ**：WARC に索引などを添え zip でまとめた、持ち運びやすい器
- ✓ WARC は**国際規格（ISO 28500）**で、IIPC が維持する
- ✓ 残した後も**長期保存（OAIS）・改変検知（fixity）**で守り続ける

最後に、Webアーカイブが人文学の研究で何に役立つかを見ておきましょう

研究資料としての Web



ソーシャルメディア、公的機関のサイト、災害の記録 — こうした**いま生まれている資料**も、残しておけば、後の世代が**研究・検証**できます

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「**数年後には消えていそう**」と感じるWebページは、どんなものですか
- ✓ それが消えたとき、**困る人**は誰でしょうか

「残す価値があるもの」を見極めることも、Webアーカイブの大切な一歩です

まとめ

- ✓ Webは**消え・変わる**ので、ある時点を丸ごと残す必要がある
- ✓ Webアーカイブは**集める・残す・再生する**の三段階
- ✓ **WARC** は、やりとり（求めと答え）ごと束ねて残す器。だから後で再生できる
- ✓ **WACZ** で持ち運びやすく、**OAIS・fixity** で長く守り続ける

いまのWebを残すことは、未来の人文学への**贈りもの**になります

出典・ライセンス

- ✓ 参照（事実確認・翻案せず）：WARC 仕様 ／ ISO 28500・IIPC（iipc.github.io/warc-specifications）
- ✓ 参照（事実確認・翻案せず）：WACZ 仕様 ／ Webrecorder（specs.webrecorder.net）

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