DH ・ 技術要素 ・ OAI-PMH

OAI-PMH入門

メタデータを集めて、横断する

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

- ✓ **クリエイティブ・コモンズ**等のオープンな資料を参照し、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**本人声のAI音声合成**
- ✓ 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

「散らばったメタデータを、機械的に集めて横断する」仕組みをつかむ

- ✓ **OAI-PMH**がどんな問題を解くプロトコルかを、自分の言葉で説明できる
- ✓ **データプロバイダ**と**サービスプロバイダ**の2つの役割を区別できる
- ✓ **6つのverb**と、レコードが**ヘッダ＋メタデータ**からなることをイメージできる
- ✓ セットや日付での**差分収集**が、横断検索・集約を支えると説明できる

前提: 特にありません。「メタデータ」「目録」という言葉に触れたことがあれば十分です

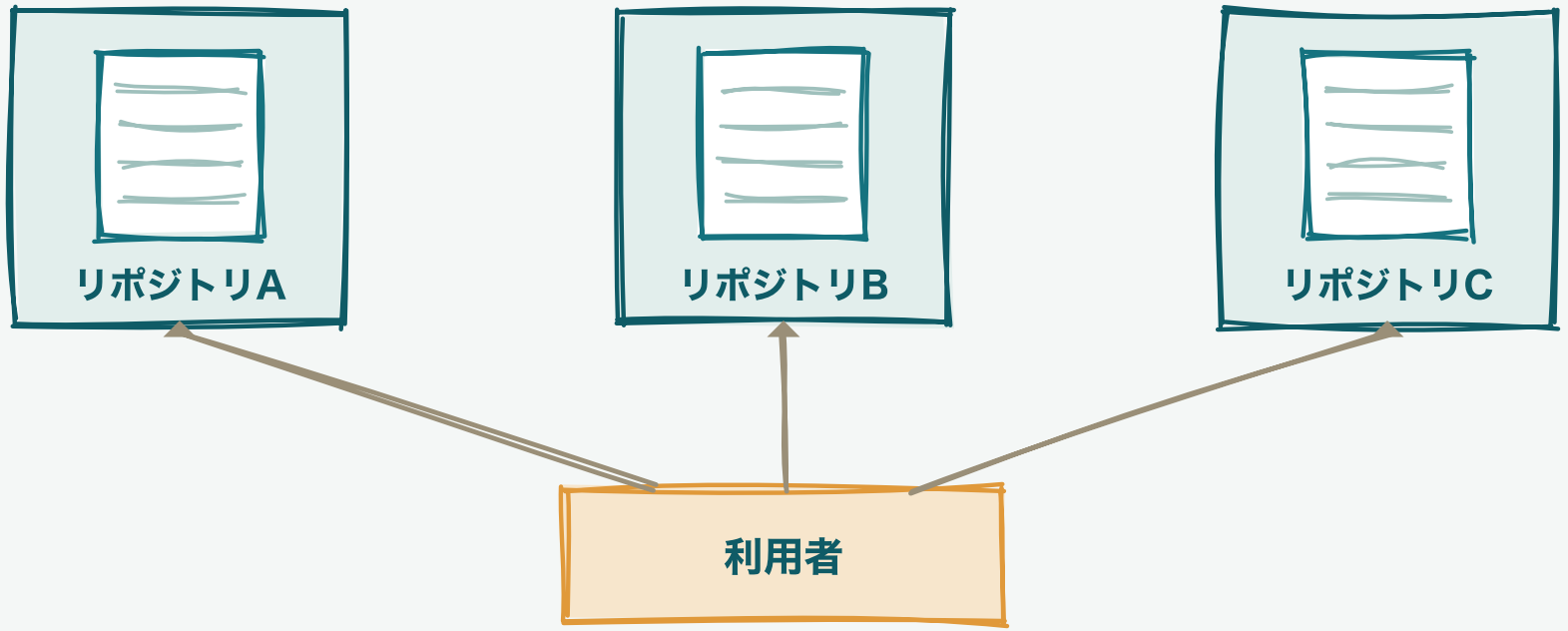
今日の流れ

- ✓ なぜメタデータを**集めたい**のか — 散らばる目録の問題
- ✓ **2つの役割** — 出す側（リポジトリ）と集める側（ハーベスタ）
- ✓ **6つのverb**とレコードの形 — HTTPで尋ね、XMLで返る
- ✓ **セットと日付**で選んで・差分で集める
- ✓ **横断検索・集約**と、IIIF・DTSとの対比、はじめの一歩

なぜメタデータを集めるのか

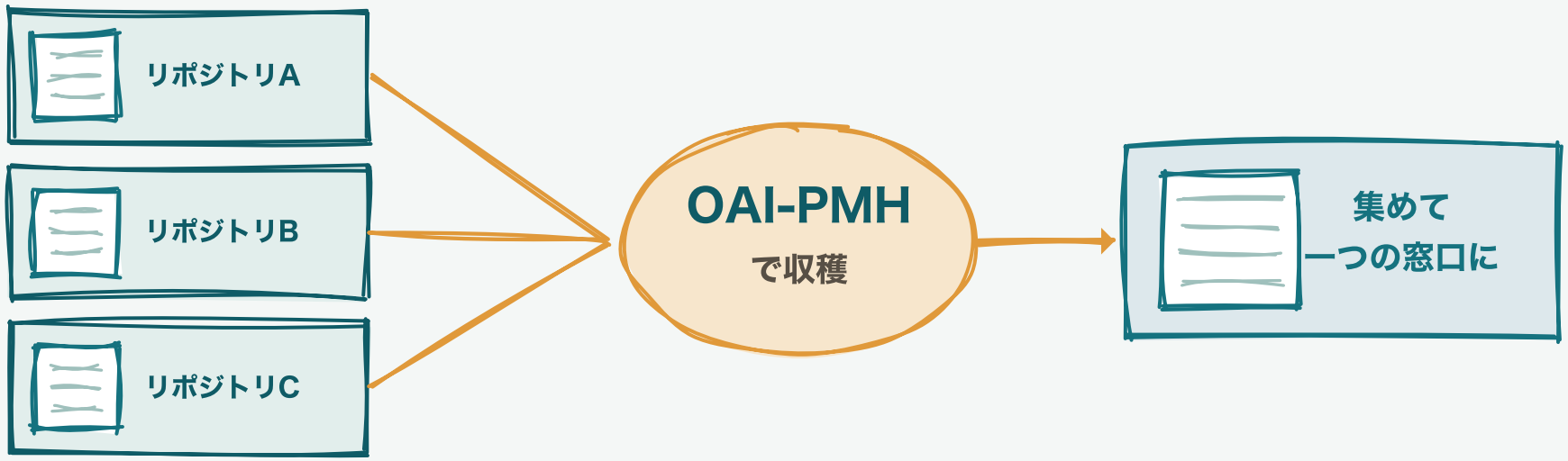
まずは、目録が「散らばっている」問題から

目録が、館ごとに散らばっている



資料の**メタデータ**（目録情報）は各リポジトリの中。横断して探すには、**一つずつ**巡るしかない

OAI-PMH ＝ メタデータを「収穫」する約束



OAI-PMHは、リポジトリから**メタデータを機械的に集める**（ハーベストする）ための、共通プロトコル

ここまでの整理

✓ 資料の**メタデータ**は、館ごとのリポジトリに散らばっていた

✓ 横断して使うには、機械的に**集める**仕組みが要る

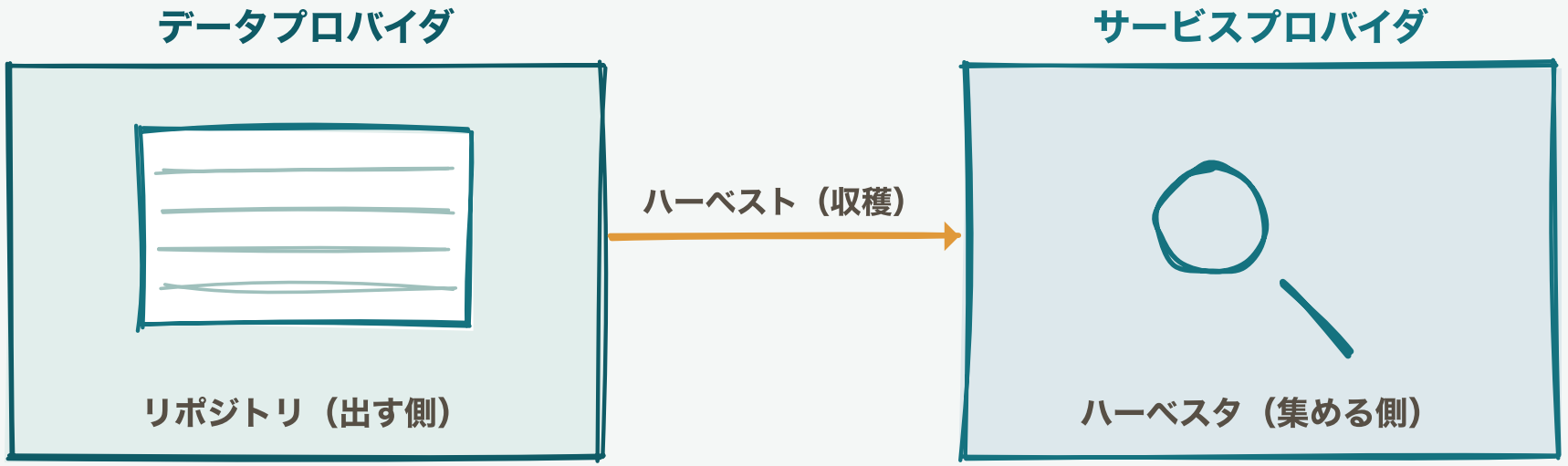
✓ **OAI-PMH**は、その「収穫」の手順を取り決めた古くからの標準

大事な前提 — OAI-PMHが運ぶのは**メタデータ**で、資料そのもの（本文や画像）ではありません

2つの役割

出す側と、集める側

データプロバイダと、サービスプロバイダ



出す側＝**データプロバイダ**（リポジトリ）、集める側＝**サービスプロバイダ**（ハーベスタ）

HTTPで尋ね、XMLで返る

やり取りは**HTTP**のURL（要求）と**XML**の応答だけ。特別な仕掛けは要らず、広く実装しやすい

ここまでの整理

- ✓ 出す側の**データプロバイダ**が、メタデータを公開する
- ✓ 集める側の**サービスプロバイダ**が、それをハーベストする
- ✓ やり取りは**HTTP**の要求と**XML**の応答 — シンプルで実装しやすい

では、ハーベスタは具体的に「何を」尋ねるのでしょう。つぎは**6つのverb**を見ます

6つのverbとレコードの形

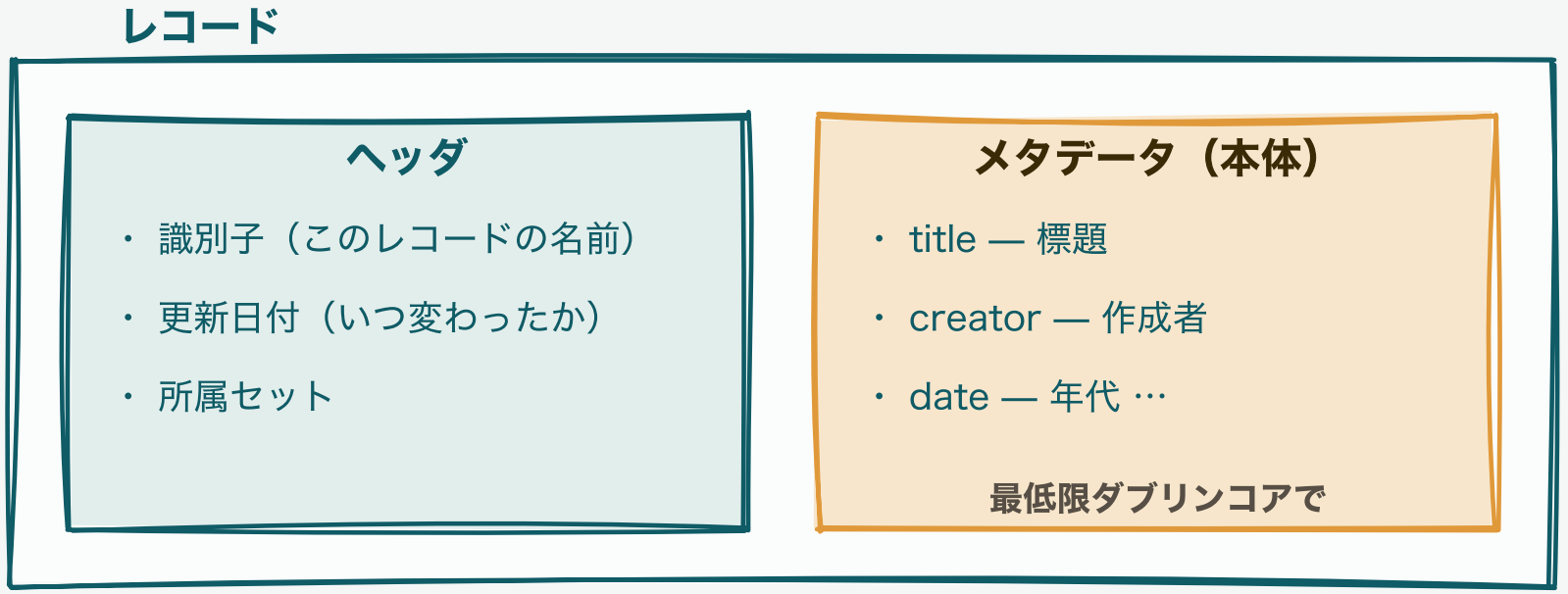
何を尋ね、何が返るのか

尋ね方は、たった6つの動詞

尋ね方は、この6つだけ

この**6つ**だけ。**ListRecords**でまとめて、**GetRecord**で一件ずつメタデータを受け取る

レコード＝ヘッダ＋メタデータ



各レコードは**ヘッダ**（識別子・更新日付・所属セット）と**メタデータ**本体。最低限**ダブリンコア**で返せる約束

選んで・差分で集める

セットで範囲を絞り、**更新日付**で前回以降だけを取る。大量データは**続きトークン**で小分けに

ここまでの整理

✓ 尋ね方は**6つのverb**。一覧で集め、必要なら一件ずつ取る

✓ レコードは**ヘッダ＋メタデータ**。最低限**ダブリンコア**で返せる

✓ **セット・日付・続きトークン**で、必要な分だけ・差分で・小分けに集められる

この仕組みがそろうと、何がうれしいのか。つぎは**集約**の話です

何がうれしいか

集めて、横断する

集約

ばらばらの目録を、一つの窓口に

多くのリポジトリのメタデータを集約し、**横断検索**できる**ポータル**を作れる。利用者は一度で見渡せる

たとえば、こんな場面で

- ✓ 多数の**機関リポジトリ**の論文・資料情報を集め、横断検索サービスにする
- ✓ 図書館・博物館の目録を集約し、**分野横断のポータル**を作る
- ✓ 集めたメタデータを**定期的に更新**（日付で差分収集）して鮮度を保つ

「各館を一つずつ見て回る」から、「一か所で見渡す」へ。集約が**発見**を助けます

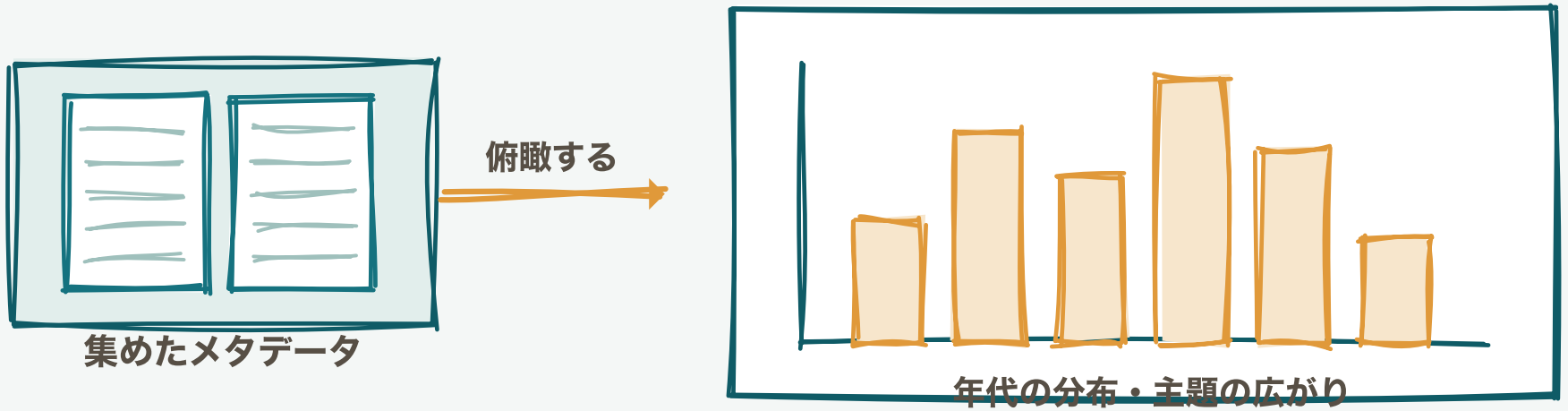
集約

「メタデータ」を集める — IIIF・DTSとの対比

それぞれ「何を」運ぶかが違う — 互いに補い合う

OAI-PMHが運ぶのは目録＝メタデータ。画像は**IIIF**、テキストは**DTS**。互いに補い合う関係です

デジタル・ヒューマニティーズでの活用



集めたメタデータは、**分析**の素材にもなる。年代の分布、主題の広がりなど、**大きく俯瞰**して問いを立てられる

考えてみよう

少し、考えてみましょう

あなたが横断して探したい資料は、どこにありますか ―

✓ いくつの**リポジトリ**に、ばらばらに置かれていますか

✓ それらを**一か所**で見渡せたら、どんな問いを立てられそうですか

よろしければ、ここで一度動画を止めて、思い浮かべてみてください

留意点

集めたあとに、残る問い

- ✓ 集約の質は、各館の**メタデータの質・粒度**に左右されます（ばらつきの調整が要る）
- ✓ 運ぶのは**メタデータ**。本体（本文・画像）は別の仕組みで辿ります
- ✓ 古くからの堅実な標準ですが、近年は**API**や ResourceSync など別の選択肢も。用途で選びます

「集められる」ことと「意味のある集約になる」ことは、別の課題として残ります

自分で触れてみるなら

- ✓ 公開されている**OAIエンドポイント**に、**?verb=Identify** を付けて開き、返るXMLを眺める
- ✓ **verb=ListRecords** に **metadataPrefix=oai_dc** を付けて、レコードの形を見る
- ✓ 体系的に学ぶなら、**OAI-PMH公式仕様**（v2.0）の各verbの説明を読む

まずは「URLにverbを付けて、返ってくるXMLを見る」。一往復で、輪郭がつかめます

まとめ

- ✓ **OAI-PMH**は、リポジトリから**メタデータを集める**（ハーベストする）共通プロトコル
- ✓ 出す**データプロバイダ**と集める**サービスプロバイダ**が、**HTTP+XML**でやり取り
- ✓ **6つのverb**、ヘッダ＋メタデータ（**ダブリンコア**）、セット・日付・続きトークンで差分収集
- ✓ だから、ばらばらの目録を**横断検索・集約**できる（運ぶのはメタデータ＝IIIF/DTSと補完）

資料を「各館に閉じた目録」から「集めて見渡せる資源」へ。そんな視点を手にできました

出典・ライセンス

本動画は、以下の公開資料を参照して作成しました。

- ✓ The Open Archives Initiative Protocol for Metadata Harvesting v2.0 ／ Open Archives Initiative — CC BY-SA 2.5

仕様（CC BY-SA）は事実確認・着想元としてのみ参照し、本文・図・例は中村による新規作成です（翻案・複製はしていません）。

ご清聴ありがとうございました