JAPAN SEARCH　国立国会図書館

ジャパンサーチにおける メタデータ連携の方針と技術

国立国会図書館　川島隆徳

目次

背景

背景：ジャパンサーチとは

- ジャパンサーチは、書籍・公文書・文化財・美術・人文学・自然史/理工学・学術資産・放送番組・映画など、我が国が保有する様々な分野のコンテンツのメタデータを検索・閲覧・活用できるプラットフォーム
- 複数の検索システム（デジタル・アーカイブ）からメタデータを収集し、検索できるようにするシステム
- 多様なデータ
  - 191データベース、2600万レコード（2022/10/12）
  - 書籍、美術、博物、地域資料からデータセットまで

背景：システムの要求事項と課題

- 連携の容易性（＝持続可能性）
  - トータルでの連携コストをミニマムに
    - システム連携コスト
    - 投入作業コスト
    - メタデータマッピングのコスト

- 最低限の項目の共通化
  - 特に、システムの特徴であるコンテンツのライセンスなど
  - 検索、画面の構成にも最低限は必要

方針

連携をシンプルにするための4つの方針

方針：（１）大がかりなスキーマを作らない

- 従来、こういった連携データベースの構築に当たっては、大がかりなメタデータスキーマを策定し、それに各データをマッピングしていくやり方が一般的（NDLサーチ、Europeana）
  - スキーマの策定そのものが目的化してしまうこともあるように見える「なぜスキーマを考えるか？それは、そこにデータがあるから」
  - RDBの名残でもあるか
- 大規模なスキーマを理解・運用することは高コスト
  - 項目の意図の継承
  - マッピングのルール
  - マッピングできない項目
- データが多様すぎてマッピングを作るのも大変

方針：（１）大がかりなスキーマを作らない

- 方針
  - オリジナルのメタデータを保持する（＝元データのマッピングを利用する）
  - 最低限の共通項目を、オリジナルのメタデータからコピーする形で持つ
  - 何もかもを綺麗に表示するのは諦める

方針：（２）簡単なシステム連携

- メタデータのシステム連携で差分更新をしようとすると、安易に辿り着くのはOAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）
- メジャーな技術では無いため、収集側は簡単だが、連携側（データ提供側）の実装が高コスト。
  - 後継のResourceSyncというのもあるが、さらにマイナー
- 現在のシステム処理能力・通信速度で、そもそも差分処理が必要か？
  - 経験上、数百万オーダーになってくると、「少し」時間がかかる

方針：（２）簡単なシステム連携

- 方針
  - ベースは全件一括取得、一括更新
    - ファイルアップロード or wget
  - 一部の大規模DB用にOAI-PMHも備える
  - 受け取る側（ジャパンサーチ側）で多様なファイル形式に対応
    - CSV/TSV/XLSX/XML/JSON
  - 連携は手動でも、自動（定期実行）でも可
  - 連携機関がDIYできるように（=システム管理側がボトルネックにならない）

方針：（３）・（４）事後処理

（３）検索時マッピング

- 緻密なマッピングを放棄すると、検索項目は貧弱になる
- 方針：検索項目を自由に定義できる仕組みを用意することで、複雑な検索要件を充足

（４）利活用マッピング

- 所在検索以上のデータの活用を行うためには、既存の語彙との紐付け、統制など、メタデータへのアノテーションが必須
- 方針：元データとは別にRDFストアを用意し、そこは単純な連携とは切り離し、少数精鋭で自由にやってもらう

仕組

データモデル

検索定義

検索ボックスを生成するための定義

データベースを管理する組織
ないしは、個々資料の所蔵組織（つなぎ役を介する場合）

組織

1回のデータ収集
収集のいずれか一つの断面が、公開されている

一つのデータベース＝連携先
データベースレベルで共通の属性
収集の設定

データベース

収集

収集
ファイル

ラベル

アイテム

検索インデックス

各データベースのメタデータ項目の定義
共通項目の定義

1件のメタデータ
1つのデータベースに所属

DIY

- 連携は、運営側は登録を承認するだけで、あとは連携機関が自分一人でできるように作ってある。
  - もちろん分からないことがあればサポートはする。
- 基本的には全ての作業はWeb管理画面から実行可能。
- 列名が入った小規模なTSVからの連携なら、（開発者なら）10分もあればデータの公開が可能

連携フロー

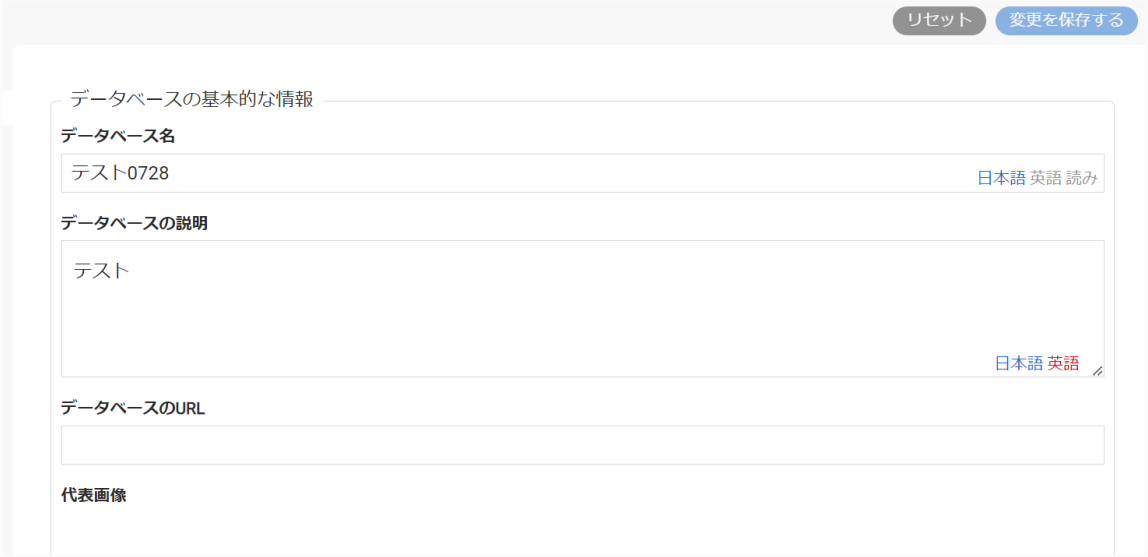
データベース登録ウィザード画面

組織登録

- データベースは特定の組織がオーナーになり、その組織に所属しているユーザが編集できる
- ジャパンサーチでは、つなぎ役と言って、複数のメタデータシステムを単一の機関が一度集約し、それが連携される、というモデルもある。
  - この場合、オーナー＝個別のアイテムの所蔵機関となるため、メタデータの項目として所蔵機関も定義出来るようになっている。
  - オーナーで無い所蔵機関については、組織を登録すると、メタデータから組織のページにリンクできる。未登録だと、ただの文字列扱い。

データベース

- 必須項目はID、名称、説明、カテゴリ
- また、データベース内のアイテムのサムネイルやコンテンツの権利区分や公開状況など、アイテムに横断的に設定する項目をここで指定することも可能。



データ投入

- 現状は以下に対応
  - 連携インタフェース：アップロード(1GBまで)、HTTPによるGET（Basic認証）、OAI-PMH
  - ファイルフォーマット：CSV/TSV/XLSX/JSON/XML/RDF
    - CSV/TSV/XLSXであれば、ヘッダをラベル定義として使える
    - JSON/XMLは、推奨は1行1JSONないしは1XML。ルート要素の子要素になっているものでも拾える
    - zip圧縮による複数ファイル対応
  - 収集頻度：ワンショット、定期実行
- 運用開始時は一部特別対応したが、最近は特に追加は無い
  - CSVやXMLでの投入が多いようである

データ分析

- 投入されたデータは、内部的には一律JSONに変換される（TSVなどは、列番号：値、といった形に）
- JSONに変換されたデータは、各項目にどういった値（文字列、数値、日付etc）が入っているかや、ユニーク数（ユニークな値の種類数）、充足数（空でないデータが入っている数）などを調べる
- 後にラベル定義や検索定義で利用する

ラベル定義

- ジャパンサーチのメタデータの構造

  https://jpsearch.go.jp/api/item/aokenshida_pic-05560_02


```
{
  "id": "aokenshida_pic-05560_02",
  "common": {
    "id": "aokenshida_pic-05560_02",
    "title": "絵ハガキ（津軽海峡の女王青函連絡船）",
    "lastUpdatedDate": 1627448027413,
    "linkUrl": "https://kenshi-archives.pref.aomori.lg.jp/il/meta_pub/G0000004pic_05560-02",
    "thumbnailUrl": [
      "https://kenshi-archives.pref.aomori.lg.jp/il/cont/01/G0000004pic/000/008/000008031.jpg"
    ],
    "contentsType": "image",
    "contentsRightsType": "ccbysa",
    "contentsAccess": "internet",
    "category": [
      "regional",
      "cultural",
      "humanities"
    ],
    "temporal": [
      "現代",
      "昭和20年代"
    ],
    "location": [
      "北海道函館市",
      "北海道函館市",
      "青森県",
      "青森市",
      "05_その他",
      "01_北海道"
    ],
    "coordinates": {
      "lat": 41.796197,
      "lon": 140.66899
    },
    "provider": "青森県",
    "ownerOrg": "prefaomori_2019",
    "database": "aokenshida_pic",
    "apiType": "ok",
    "subCategory": [
      "絵はがき",
      "青森県",
      "写真"
    ],
    "access": "PUBLIC",
    "dclass": "683"
  },
  "rdfindex": {
    "type": [
      "絵葉書"
    ],
    "spatial": [
      "日本 > 北海道"
    ]
  },
  "aokenshida_pic-12-s": "01_北海道",
  "aokenshida_pic-35-s": "CC-BY-SA",
  "aokenshida_pic-13-s": "北海道函館市",
  "aokenshida_pic-34-s": "画像の改変と営利目的を含み利用が可
  "aokenshida_pic-10-s": "昭和20年代",
  "aokenshida_pic-33-s": "函館湾　函館港　洞爺丸",
  "aokenshida_pic-11-s": "05_その他",
  "aokenshida_pic-5-s": "02_絵はがき",
  "aokenshida_pic-6-s": "05_近現代",
  "aokenshida_pic-8-s": "現代",
  "aokenshida_pic-1-s": "00-O-5560",
  "aokenshida_pic-2-s": "絵ハガキ（津軽海峡の女王青函連絡船）
  "aokenshida_pic-3-s": "洞爺丸（青函連絡船）",
  "aokenshida_pic-4-s": "1",
  "aokenshida_pic-28-s": "長島1-1-1",
  "aokenshida_pic-27-s": "青森市",
  "aokenshida_pic-26-s": "青森県",
  "aokenshida_pic-25-s": "00-O-5560",
  "aokenshida_pic-24-s": "青森県",
  "aokenshida_pic-22-s": "状",
  "aokenshida_pic-0-s": "05560-02",
  "aokenshida_pic-39-u": "https://kenshi-archives.pref.aom
  "aokenshida_pic-38-u": "https://kenshi-archives.pref.aom
  "aokenshida_pic-16-s": "41.796197",
  "aokenshida_pic-17-s": "140.66899",
  "aokenshida_pic-14-s": "北海道函館市",
  "aokenshida_pic-37-s": "2019/2/8"
}
```


共通項目

個別項目

ラベル定義

- 個別項目と共通項目がある
  - 個別項目は、オリジナルの項目に対して、その項目名や種別を定義したもの
  - 共通項目は、ジャパンサーチの共通的な項目（スキーマ）に対して、既存のデータのどの項目が対応するかマッピングしたもの
  - 共通項目は現在21項目、うち必須2項目（IDとタイトル）、データベース単位で可能な設定が7項目

https://jpsearch.go.jp/api/database/aokenshida_pic/label

個別項目ラベル

パス	項目ラベル(日)	項目ラベル(英)	格納種別	データ種別	項目説明(日)	項目説明(英)	項目例	サンプル	充足数	ユニーク値	削除
0	ID		通常	文字列			07048-08	07048-08	8,130	8,130	
1	資料番号		通常	文字列			2306	2306	8,090	3,445	
2	資料名1		通常	文字列			初冬の岩木山	初冬の岩木山	8,130	1,830	
3	資料名2		通常	文字列			(国立公園十和	(国立公園十和田湖)明...	5,201	3,781	
4	員数		通常	文字列			77	77	7,696	7	
5	種別1		通常	文字列			01_写真	01_写真	8,130	3	
6	種別2		通常	文字列			05_近現代	05_近現代	8,130	2	
8	時代		通常	文字列			近代	近代	8,116	2	
9	西暦		通常	文字列			1976	1976	3,599	104	
10	和暦		通常	文字列			昭和36年11月2	昭和36年11月2日	7,609	720	
11	地域名1		通常	文字列			05_その他	05_その他	7,997	6	
12	地域名2		通常	文字列			13_東京都	13_東京都	459	30	

共通項目ラベル

ID	必須	ID
名称/タイトル	必須	資料名1
名称/タイトル英語	あれば必須	
名称/タイトルヨミ	あれば必須	
最終更新日	あれば必須	
URL	あれば必須	データURL
サムネイル画像URL	あれば必須	サムネイル表示URL +
コンテンツURL	任意	+
IIIFマニフェストURL	あれば必須	
コンテンツフォーマット	任意	
コンテンツの権利区分	あれば必須	license
コンテンツ公開状況	あれば必須	コンテンツ公開状況
カテゴリー	任意	+
解説	任意	

ラベル定義

- 基本的に、検索結果に表示されるのは共通項目
- 詳細ページに行くと、個別のDBの項目が表示される

https://jpsearch.go.jp/item/aokenshida_pic-05560_02

共通項目

個別項目

テスト

- テスト用の検索インデックスを本番とは別に作ることができる。ここでアップロードしたデータを実際に検索・表示させてテストが可能。

テスト公開可能なデータ

公開中	名前	収集日	収集方法	フォーマット	状態	
	定期収集-20220318	2022/03/18 01:00:21	OAI-PMH	XML	一般公開処理に成功しました	テスト公開する
✔	gunma	2020/07/20 10:56:56	OAI-PMH	XML	テスト公開処理に成功しました	テスト公開する

公開

- 公開すると、データベースが公開状態となり、同時に検索インデックスの作成が動く。
  - データベースが大きいと時間がかかるので、検索件数が一定しなくなる。
  - 速度は数百レコード/秒なので、大規模なもので無ければすぐ終わる

検索インデックスの状態

- テキストを全て拾い集めてくるキーワードフィールド一つ
- 共通項目を検索できるフィールド約30
- ソートやファセットのために裏で正規化を施した検索用フィールドが10程度（表示には使わない）
- これらに加えて、登録された全てのデータベース項目が原則検索可能であり、トータルで5600程度。
  - https://jpsearch.go.jp/api/database/search-field?jps-act=M
- バックエンドの検索エンジンである、ElasticsearchのDynamic Mapping機能を使って実現している
  - https://www.elastic.co/guide/en/elasticsearch/reference/current/dynamic-mapping.html

検索定義

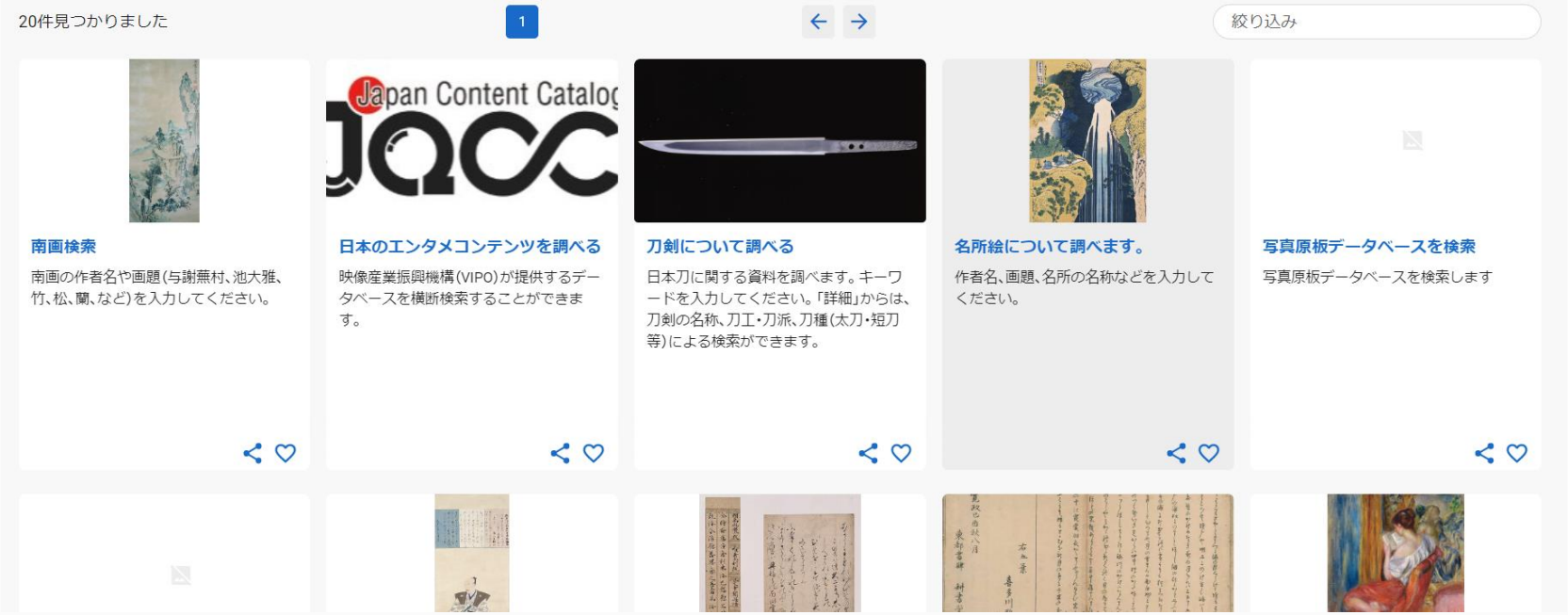
- 検索定義は、検索画面を生成するための定義
- どの項目を、どういった見た目で検索するかを定義可能

https://jpsearch.go.jp/api/csearch/jps-cross


```
},
{
  "key": "loc",
  "name": {
    "ja": "場所",
    "en": "Location"
  },
  "type": "location"
},
{
  "key": "contributor",
  "name": {
    "ja": "人物/団体",
    "en": "Contributor"
  },
  "type": "keyword",
  "fields": [
    "common.contributor"
  ]
},
```

テーマ別検索

- 検索定義の機能を使って、様々な検索画面を提供している。
- また、個別のギャラリーの中に組み込むことも可能



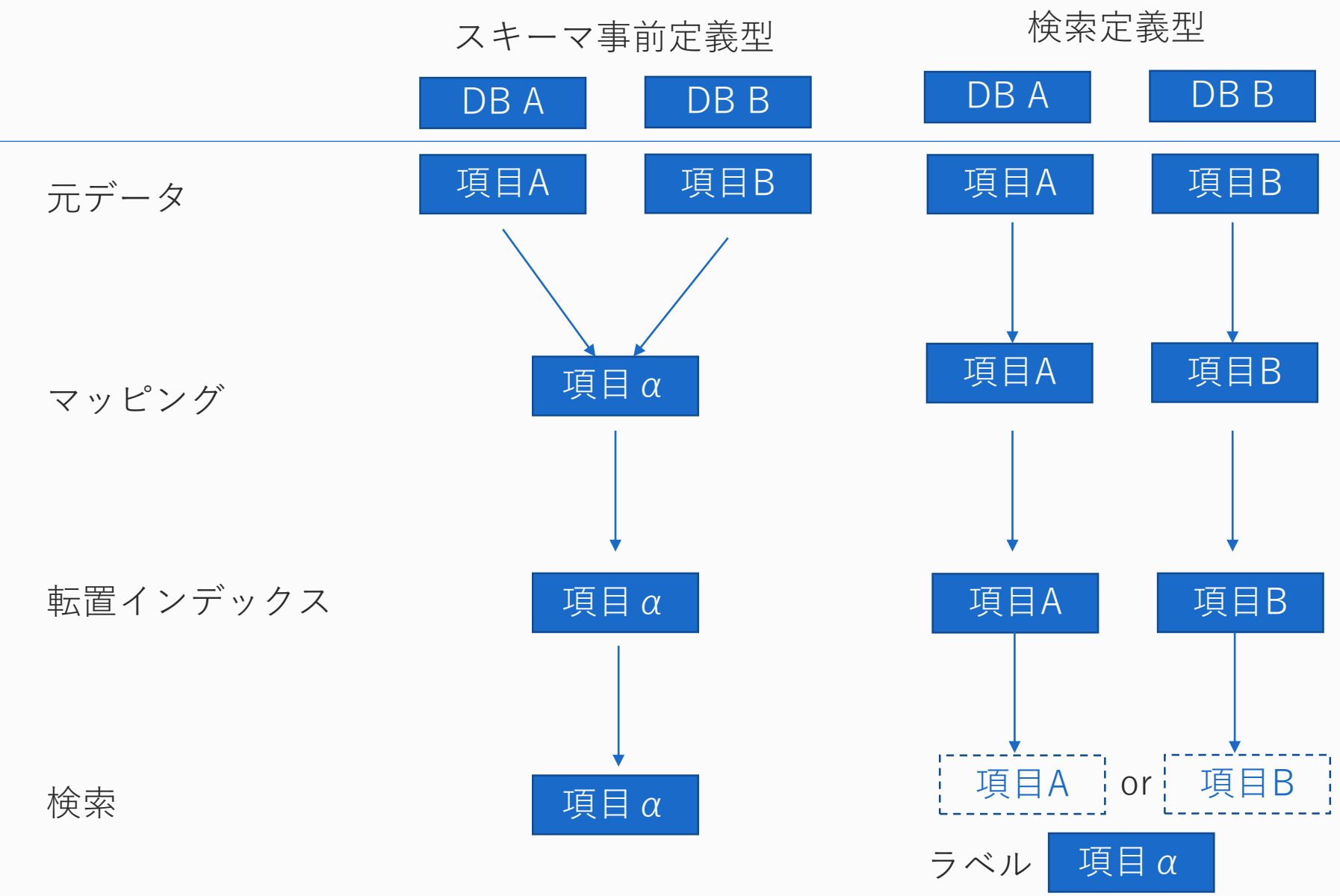
検索のカスタマイズ

- 既存の検索に、検索項目を自分で追加することも可能

検索項目を検索して、複数を選び

その項目で検索できる

横断検索システムの検索時マッピング



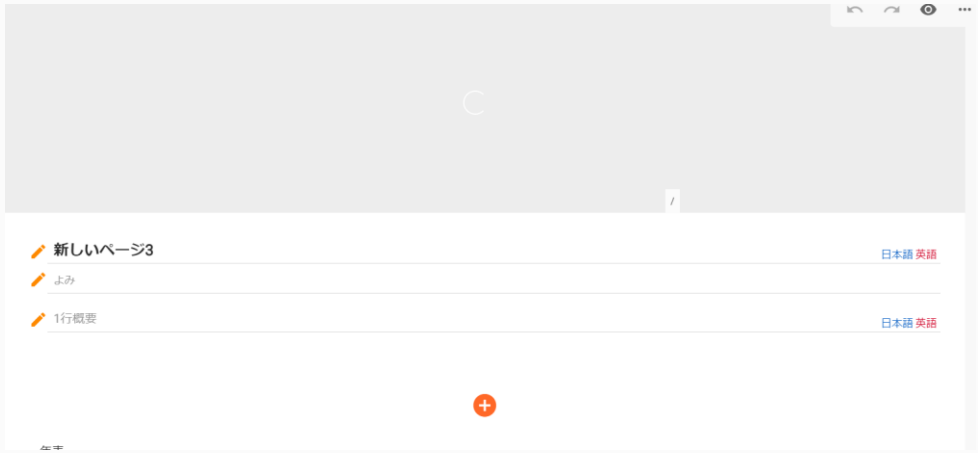
- ナイーブには等価
- 実際には、正規化、ランキングロジック（BM25）等で差異はある。
- 検索定義型はオンデマンドなので、柔軟性は高い。

利活用マッピング

- RDF化して、RDFのエンドポイントを公開
- RDF化は、提供機関の関与は無く、運用側の判断で行われている。
- RDF化のポイントは、どこまで正規化し、どこまで世界に通用する語彙にできるか、ということ。これは専門の人がバランスを見ながらやる他は無い。
- 詳細はhttps://jpsearch.go.jp/static/developer/introduction/

ギャラリー

- ジャパンサーチでは検索だけでなく、人手でキュレーションしたギャラリーも公開している。
- ギャラリーはエディタも開発しており、個人でも試すことができる。



考察

メタデータスキーマの階梯

階梯	スキーマ	中心となる主体	課題
整理	整理の段階では、整理者どのように事物を捉えているかに応じて、メタデータが決まる	整理者	• 整理者がさらに先のレイヤーを見据えてスキーマを決めることは難しい • 独創性が出てしまう
専門利用	同じ領域の専門家は、整理のスキーマを理解可能	同じ領域のユーザ	• スキーマを理解可能だからといって、検索で使いこなす（必要がある）とは限らない
（ギャップ）			
ライトな利用	検索エンジンに引っかかればよい＝スキーマレスでも良い	領域外のユーザ	• （課題ではないが）スキーマでは無く、キーワードの工夫やその他の集合知的技術により課題を解決する
リンクトデータ	世界のスキーマに従わないとリンクしない	リンクトデータ総体	• 世界のスキーマを理解できる人は少ない

メタデータスキーマの課題

- 整理と利用のレイヤー間にもギャップがある
  - 整理している人の「気持ち」「判断」をユーザは知ることができない。
  - 整理者から、利用者に向けた使い方の説明が必須。
  - 整理者は、必ずしも利用のことを考えて整理していない。

- 専門利用とライトな利用の間を埋める必要はあるのか？
  - これは、立場上あると言わねばならない
  - Googleもウェブサイト構造化で取り組んでいるとも言える

- 本当は整理の段階から、誰もが参照できる標準があったほうが良いのでは無いか？
  - あったほうが良いが、現実解にならないものはあってもしょうが無い

ジャパンサーチのシステムとしての回答

- 整理と他のレイヤーの間にすらギャップができるのに、横断検索システムでさらに重厚なスキーマのレイヤーを足しても、利用が便利になるとは限らない。
- 専門家とライトな利用の間を埋めるために、テーマ別検索や、人力検索の一種であるギャラリーを用意
- 必須では無いが、各連携元に整備をナッジする共通項目を設け、例えばライセンス、サムネイルのような政策的に重要な項目を標準化する

利用の実態

- 2022/9/1～9/30のアクセスログから(botなどを除いていない)
- 検索の92.3%は横断検索
  - 教育・商用利用検索（4.5%）や、インターネット公開資料検索(2.6%)も使われてはいる
- キーワード以外の検索の実行は全体の18.7%。ただしファセット絞り込みも含む。
  - 多いのはコンテンツ：5.1%、利用条件：3.3%、人物・団体2.7%、画像検索2.2%、時代2.1%、データベース1.6%、種類1.3%など
- ギャラリーへのアクセスは検索の77.5%

ジャパンサーチのシステムとしての課題

- システムとしての自由度の反面、機能も目的も見栄えも、ふわっとしてしまう。
  - 汎用的だが誰も使わないサービスという落とし穴が有り得る
  - 切り出されたGoogle的なもの？コンセプトのコアになっているEuropeanaはアンチGoogleがスタート地点なので、そういうことは有り得る。
  - 取り扱っているものは有象無象ではなくオーソリティではあるので、Webにオーソリティが埋没しないための取組（オーソリティであることの価値づけのロジックを、資本主義のブラックボックスに全面依存しないようにするための仕組み）、とも言えるが、そのあたりは検索エンジンと市場がやってくれる説もある。

サービスとしての取組はこちら：
https://jpsearch.go.jp/about/actionplan2021-2025