ジャパンサーチ（試験版）のドメイン設計思想

国立国会図書館 川島隆徳

2019年10月20日
日本図書館情報学会 第67回（2019年度）研究大会 シンポジウム

ジャパンサーチとは

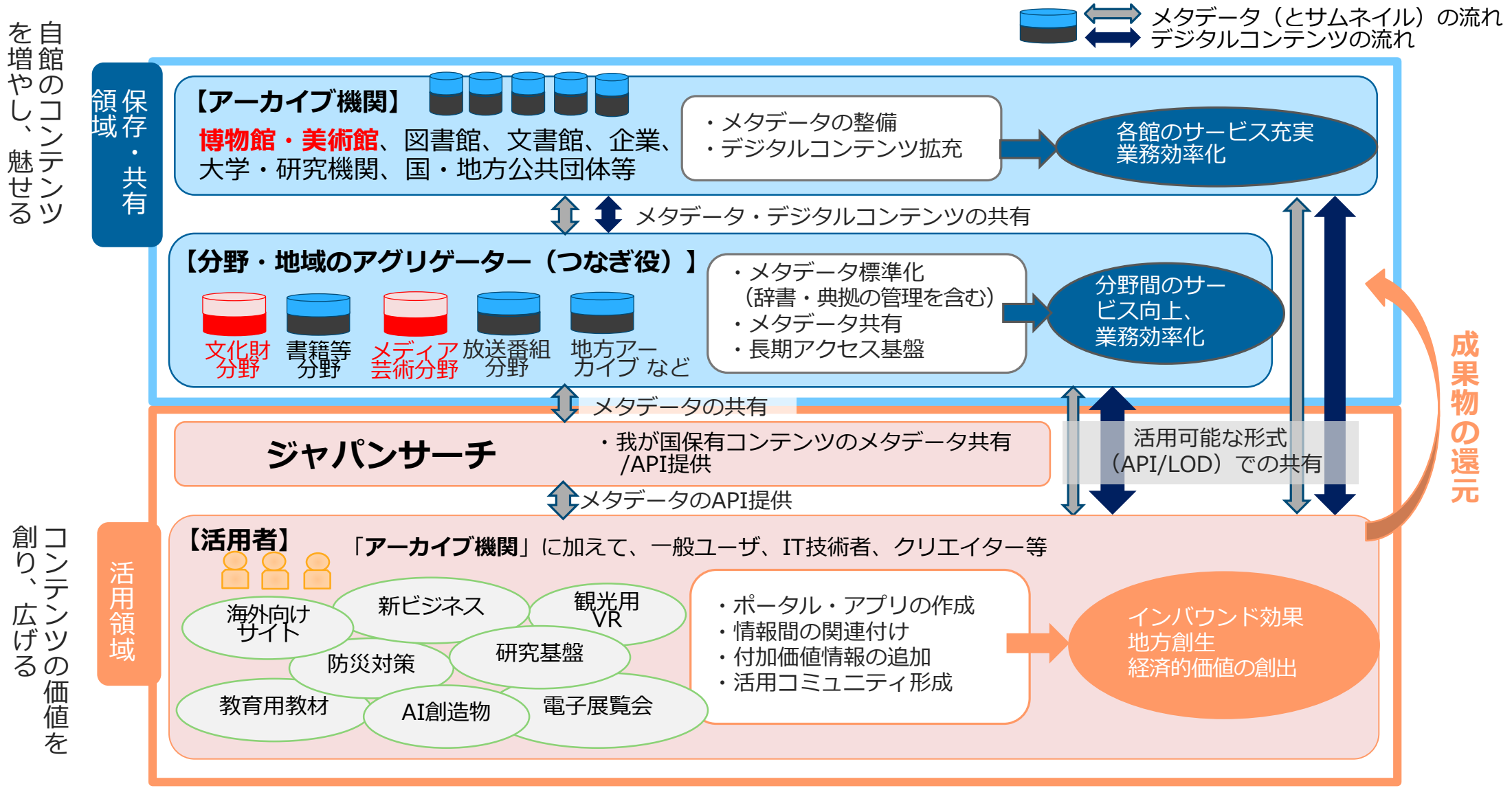
- さまざまな分野のデジタルアーカイブと連携し、我が国が保有する多様なコンテンツの**メタデータ**をまとめて検索できる**国の分野横断統合ポータル**

- **政府の「知的財産推進計画」等に掲げられている国の取組**
  運用主体：デジタルアーカイブジャパン推進委員会・
  実務者検討委員会
  （事務局：内閣府知的財産戦略推進事務局）
  システムの運用担当：国立国会図書館

- **2019年2月に試験版を公開**
  2020年に正式版の公開を目指す

ジャパンサーチ（試験版）

エコシステム



出典：デジタルアーカイブの構築・共有・活用ガイドライン（デジタルアーカイブの連携に関する 関係省庁等連絡会・実務者協議会、平成29年４月）
「図2 デジタルアーカイブの共有と活用のために」より一部修正

連携状況（2019年10月10日現在）

14機関　50データベース　メタデータ約1,800万件

分野	データ提供機関	データベース名
書籍等	国立国会図書館	「国立国会図書館サーチ」から、3件のデータベース
公文書	国立公文書館	「国立公文書館デジタルアーカイブ」
文化財	文化庁	「文化遺産オンライン」から、国指定文化財等データベース
	国立文化財機構	「ColBase 国立博物館所蔵品統合検索システム」
美術	国立美術館	「国立美術館所蔵作品総合目録検索システム」
		「アートコモンズ」
	日本写真保存センター	「写真原板データベース」
メディア芸術	映像産業振興機構	「Japan Content Catalog」から、2件のデータベース
舞台芸術	早稲田大学坪内博士記念演劇博物館	「演劇情報総合データベース」から、2件のデータベース
自然史・理工学	国立科学博物館	「サイエンスミュージアムネット S-Net」
		「魚類写真資料データベース」
人文学	人間文化研究機構	「人間文化研究機構統合検索システム nihuINT」から、30件のデータベース
	立命館大学アート・リサーチセンター	「ARC浮世絵ポータルデータベース」「ARC古典籍ポータルデータベース」
放送番組	放送番組センター	「放送ライブラリー公開番組データベース」※ドラマのデータ
	日本放送協会	「動画で見るニッポンみちしる」
公共データ	総務省・内閣府IT総合戦略室	「データカタログサイト」

- 実務者検討委員会の連携方針*に基づき、同委員会の承認を得て連携

*「第二次中間取りまとめ」（実務者検討委員会，平成31年4月）p. 33

ジャパンサーチの役割と機能要件

探す

探している人がコンテンツにたどり着けるようにする

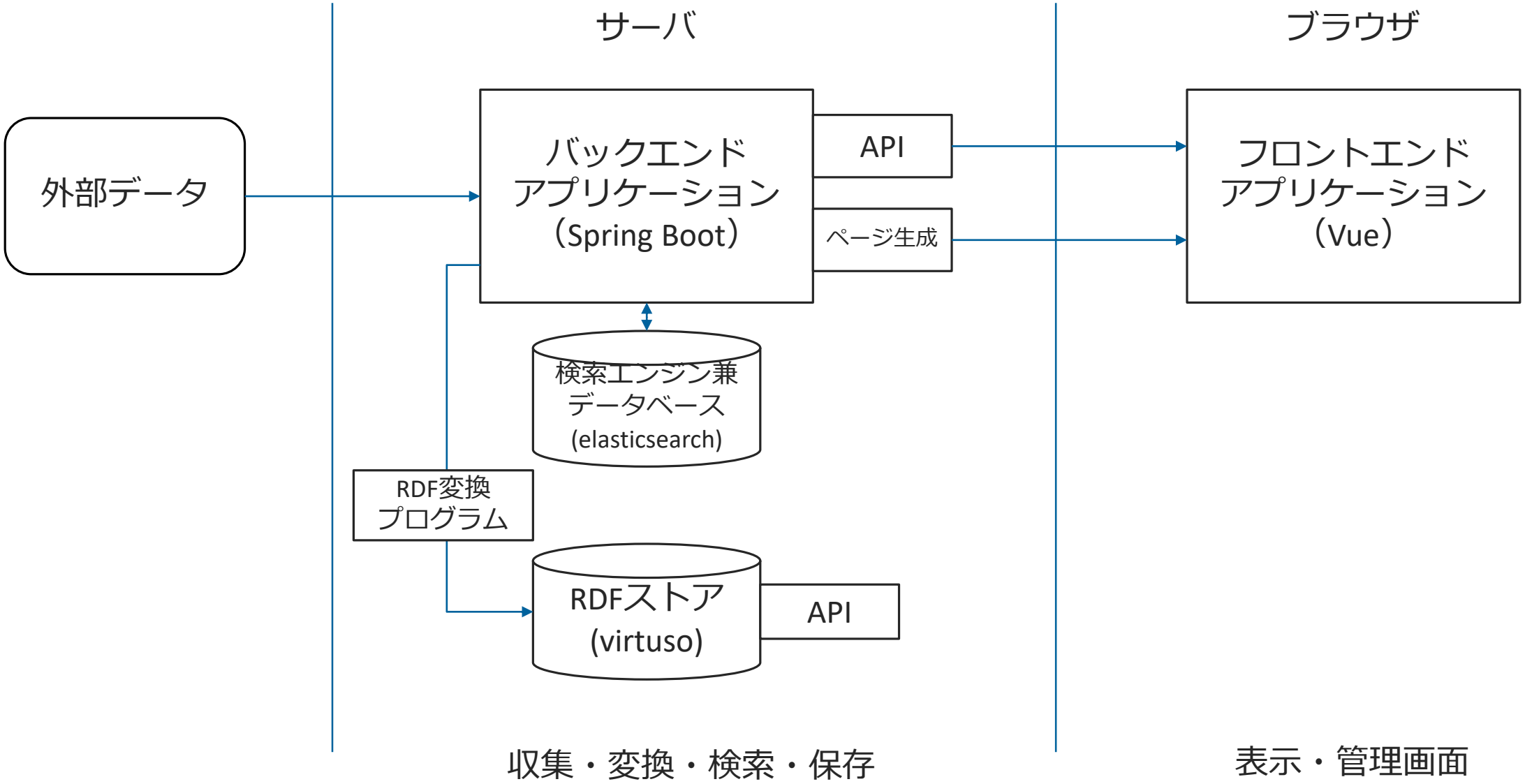
活かす

たどり着いた人がコンテンツを使えるようにする

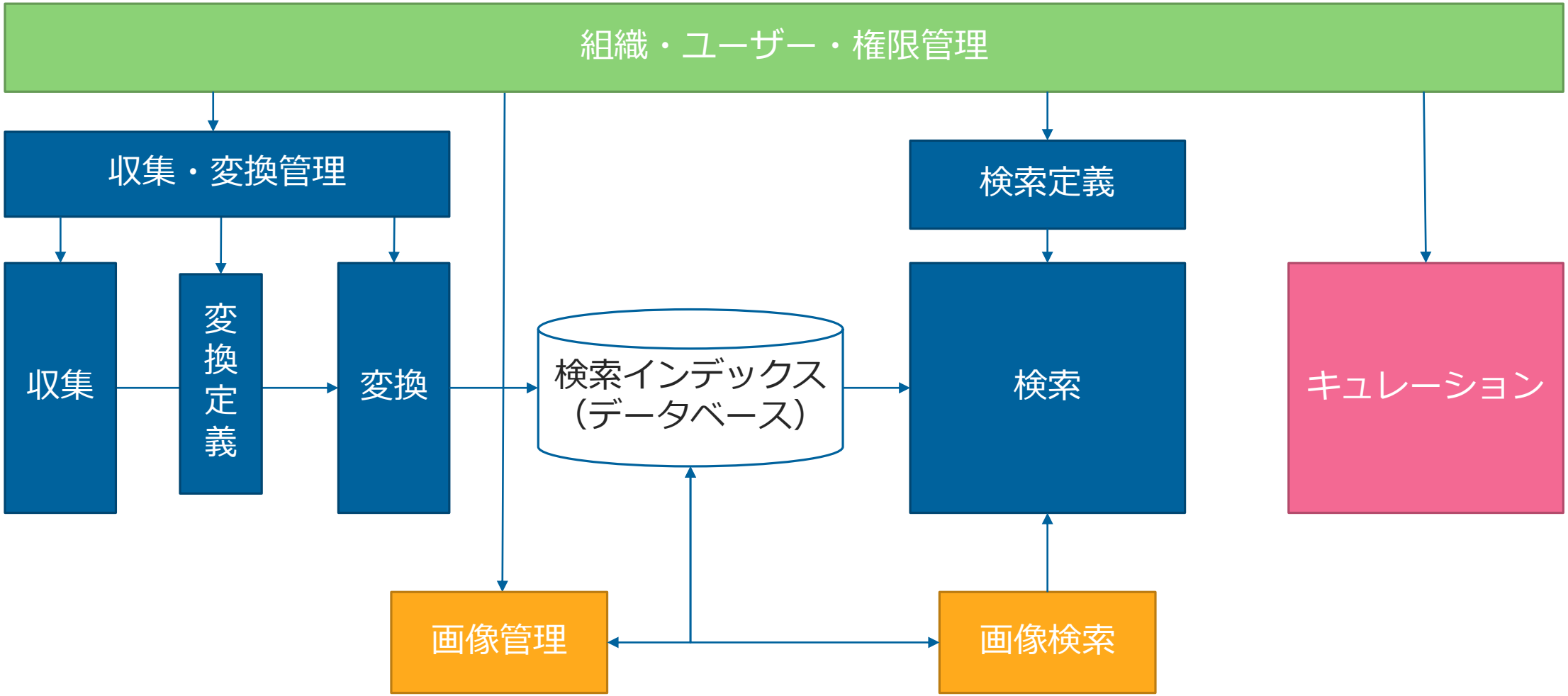
※ターゲット：専門家を含むなるべく多くの人

1.全体像：ジャパンサーチ（試験版）のアーキテクチャ

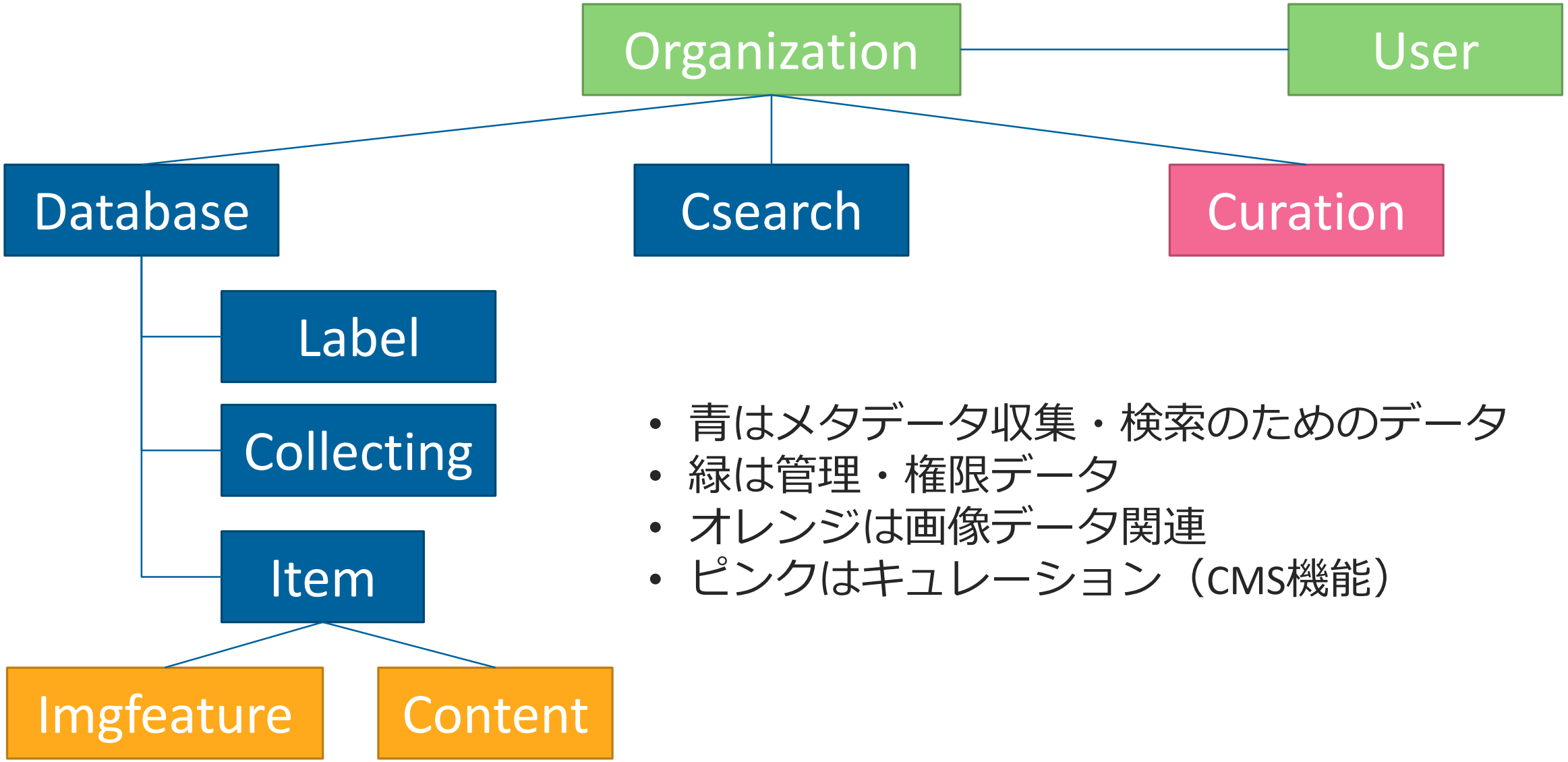
システム構成


サーバ
ブラウザ
外部データ
バックエンド
アプリケーション
（Spring Boot）
API
ページ生成
フロントエンド
アプリケーション
（Vue）
検索エンジン兼
データベース
(elasticsearch)
RDF変換
プログラム
RDFストア
(virtuso)
API
収集・変換・検索・保存
表示・管理画面

バックエンドアプリケーション構成



ドメイン（データ）設計



- 青はメタデータ収集・検索のためのデータ
- 緑は管理・権限データ
- オレンジは画像データ関連
- ピンクはキュレーション（CMS機能）

2.収集と検索：Database, Label, Item, Csearch

収集と検索の前提と方針

収集の前提
- データベース単位でメタデータを収集する。
- メタデータのフォーマット（物理・論理）は不統一。
- 連携調整コストが高いと、肝心のデータが集まらない。

⇒シンプルなマッピングで、連携のハードルを下げる＆究極のメタデータフォーマットに関する論争を回避。

検索の前提
- ユースケースが定まらないので、いろいろな検索をしたい可能性がある。

⇒この際、いろいろな検索ができるように検索そのものを抽象化して定義したい。

収集：DatabaseとCollecting

Database

データの管理の単位。名称・説明の他、データベース単位でのライセンス表示等のメタデータを持つ。

Collecting

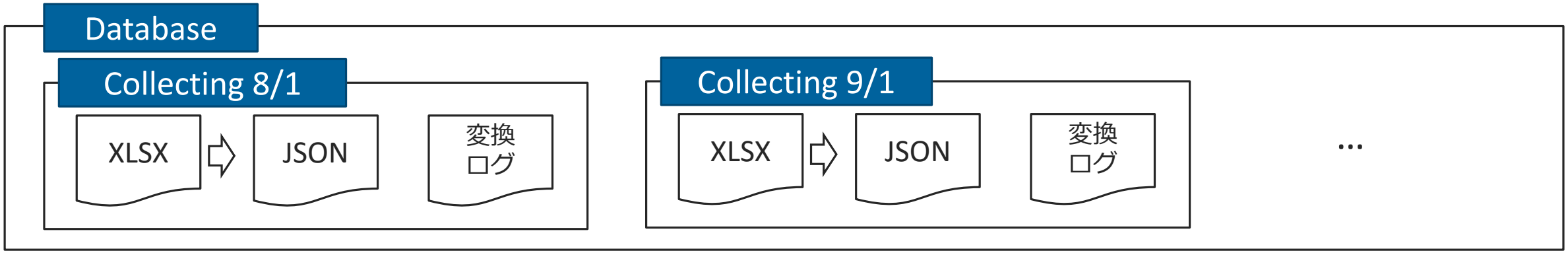
データベースからの1回分のデータ収集（断面）を示す。収集方法、変換方法、収集日等を保持。

収集方法

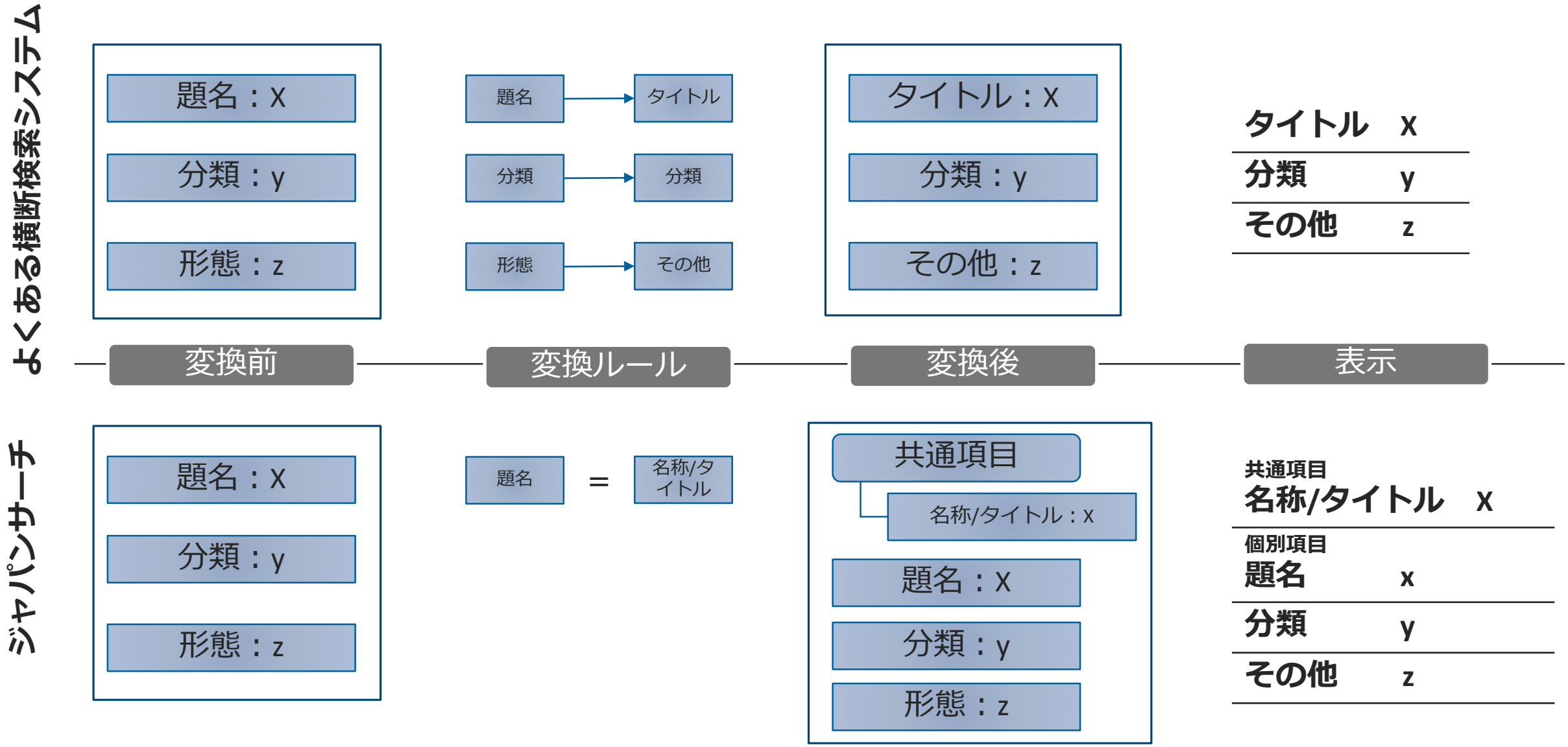
ファイルアップロード、ファイル収集、OAI-PMHによる取得等（シングルショット/定期収集）

変換方法

変換方法とは、相手のファイルフォーマット（XLSX、TSV、XMLetc）からJSONへの変換方式を指す。JPS内では、メタデータはすべてJSONで扱われる。



メタデータマッピングに関するジャパンサーチの特色



オリジナルデータを完全に保持+共通項目マッピングで最低限の共通項目を抽出（コピー）

Label

- Label
  - Field Label
  - Common Label

あるデータベースの各項目の中身の説明/定義

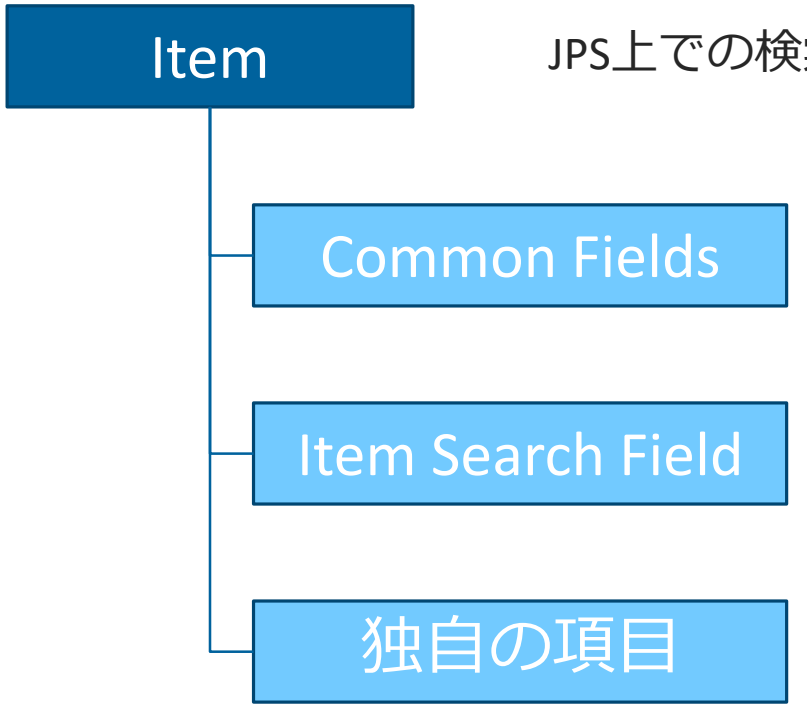
各項目のラベル名、説明等を記述する。個別データベースの項目は、この定義に従って画面上で表示される。また、表示有無や、インデックス作成方法も指定する。

どの項目が共通項目に該当するかを示す。

Field Labelの編集画面

Common Labelの編集画面

Item



JPS上での検索の単位

Common Labelでマッピング（コピーされた）共通項目が入る場所

正規化した時間や座標、並び替え用のキー等、オリジナルのメタデータにはないが、システムで検索用に付与した情報が入る。

個別のデータベース由来の各項目がそのまま格納される。

Elasticsearchには、この構造で投入される。

Relational Databaseとelasticsearch

Relational Database （RDB）

ID	タイトル	分類	その他
1	X	Y	Z
2	A	B,C	
3	P		Q

- 列（テーブル）を定義して、1行1データで値を入れていく。
- 1対多対応の構造がある場合には、テーブルを分ける必要がある場合も。
- データ構造を決めないと、システムが作れない。

elasticsearch

```
ID:1
{
  title:X,
  category:[Y],
  extent:Z
}
```

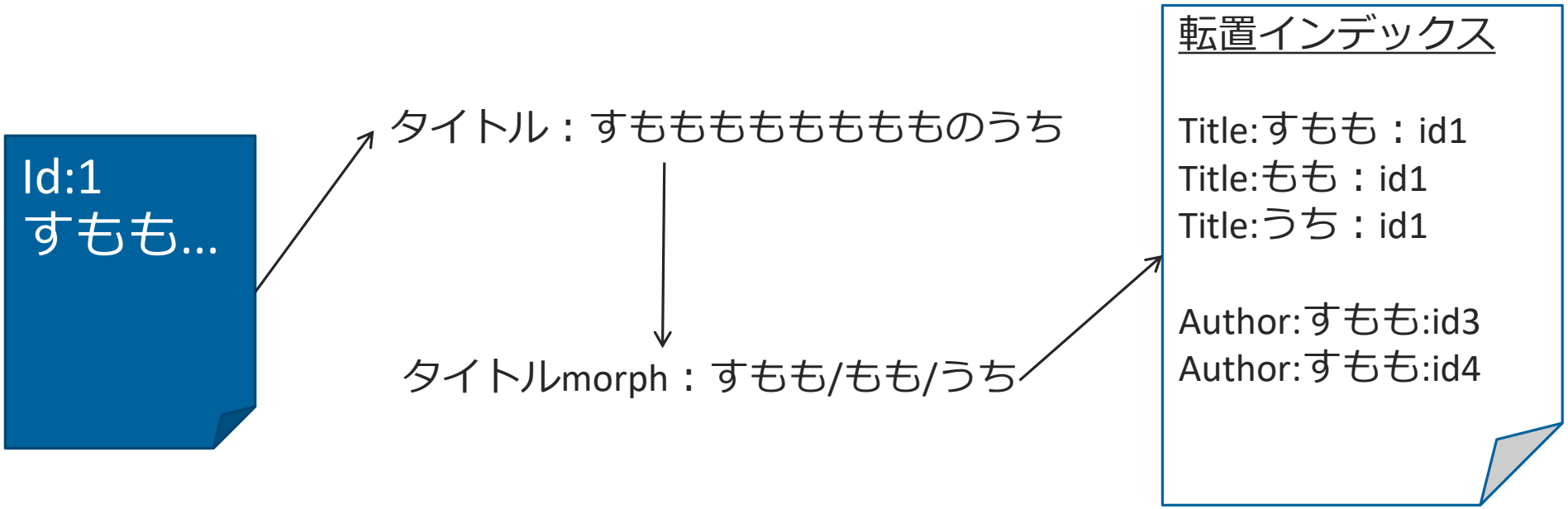
```
ID:2
{
  title:A,
  category:[B,C]
}
```

```
ID:3
{
  title:P,
  extent:Q
}
```

- Elasticsearchはdocument store型のデータベースエンジンで、任意の構造のJSONを保存し、ID指定で取得することができる。
  - 事前に構造を決めなくても良い
- 検索のためには、検索用の転置インデックスが生成される。

転置インデックス

どのドキュメントに、どの単語が登場するか　－転置→　特定の単語が登場するドキュメント



	1	2	3	4	5	6	7	8	9	10
Author:すもも	○		○		○			○		○
Title:すもも		○	○	○		○		○		
AND検索			○					○		
OR検索	○	○	○	○	○	○		○		○

→ 著者とタイトルにすももを含む資料

→ 著者かタイトルにすももを含む資料

JPSのインデックス

Database: sample

```
ID:1
{
  common:{
    title:X
  },
  title:X,
  category:[Y],
  extent:Z
}
```

common.title:X : id1
sample.title:X : id1
sample.category:Y : id1
sample.extent:Z : id1

全データベース共通のインデックス

データベース固有のインデックス

common.title:x
sample.category:y
test.category:y

1	2	3	4	5	6	7	8	9	10
○	○	○		○		○	○		○
○		○		○					
						○	○		○

→ 共通項目なら、全DB横断で検索できる

→ 個別項目なら、sample DBだけ検索

→ 複数のDBの異なる項目をOR検索すると、疑似的に共通項目検索ができる。

メタデータ検索とは

1. 特定のフィールドに特定の文字列/数値/日付etcが入っている
2. 特定のフィールドに特定の文字列/数値/日付etcが入っていない（NOT）
3. 1や2の組み合わせ（AND/OR）
4. 「特定のフィールド」が何なのか、利用者が分かる
5. 特定の条件（利用者が制御できない条件）を持つこともある
   e.g. 利用制限のかかっている資料が一般ユーザの検索対象外になる

検索（画面）を生成するための定義を自由に書ければ、
あらかじめメタデータを統合する必要は必ずしもない

欠点：異なるIndexのOR検索では、検索の高速化やランキングのチューニングが難しくなる。

Csearch=Custom Search（=テーマ別検索）

- **Csearch**：JPS上で実行可能な検索を定義する。名前の他、対象データベース等
  - **Field**：検索項目。項目名の他、どのインデックスを検索するか。
  - **Facet**：ファセットの定義。どの項目でファセットを提示するか。
  - **Column**：テーブル表示の定義。どの項目をどのようにマッピングして表形式で表示するか。
  - **Constraint**：特定の制約条件で、検索範囲をあらかじめ絞る

Csearchを定義すると、対応した検索画面が自動生成される。
専門家の使用に適した詳細な検索画面を自由に作ることができる。

※構造をもったクエリでの検索はこれではできないので、利活用フォーマットのSPARQLに任せている。

Csearch編集画面

1.XLSXファイル（例）

c1	c2
1	X
2	Y

sample データベース

2.JSON

{c1:1, c2:X}
{c1:2, c2:Y}

Common Label

ID:c1
Title:c2

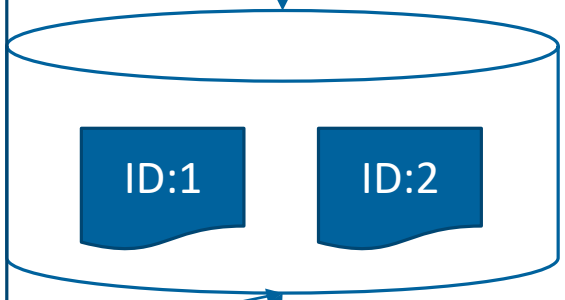
3.共通項目付きJSON

{common:{id:1,title:X}, c1:1, c2:X}
{common:{id:2,title:Y}, c1:2, c2:Y}

検索インデックス

common.title:X
common.title:Y
sample.c2:X
sample.c2:Y

Sample-1	Sample-2
○	○
○	
	○



Field Label

c1:識別子
c2:題名

詳細表示

Custom Search

ID:c1
タイトル：c2

ID
タイトル

Custom Search

ID:common.id
タイトル：common.title

ID
タイトル

共通項目
名称/タイトル　x

個別項目
識別子　1

題名　x

理想と課題

理想

- その道の専門家が、特定のドメインを検索するためのCsearchを定義して、初心者の検索をやりやすくする
  （＝検索ノウハウのシステム化）
- 柔軟なフレームワークを手にして、様々な要求に対応可能にする

課題

- 機能が複雑すぎて使ってもらえない（UIの洗練で解決可能か）
- そもそも柔軟な検索は必要とされているか？
  →様々なメタデータフレームワークやスキーマの検討の目的は？
  →検索目的ではなく表現が目的？目的の無い表現とは？
- 詳細表示などをすると画面の統一感が無い

3.広い意味での検索：Curation

検索とは？

- 検索とはGoogle？
  - ナレッジパネル
- キーワード検索をするのか？
- 受動型の情報探索
- スマートスピーカ、スマートアシスタント？

メタデータを検索できるようにするだけでなく、理解しやすい、利用しやすい形でまとめておく必要があるのではないか。

Curation（＝ギャラリー）

Curation　メタデータ、テキスト、画像等のパーツを組み合わせた任意のまとめページ

パーツ一覧

リスト	手動選択ないしは検索条件指定でメタデータのリストを表示できる。 表示方式は、スライド・パネル・表、等様々な形態がある。
画像	画像とその解説を書ける
検索BOX	任意の検索の検索窓を置ける
地図	リストを地図として表示するパーツ
セクション	セクション構造
テキスト	任意のテキスト
サブページ	複数ページの表現

普通のCMSとの違い

- 出来ることはほぼ普通のCMS（Word Press等）と変わらない
- メタデータのリストなどを、ベタうちではなくて構造化情報として表現できるのが強み
  →画面に表示する以外の再利用も相対的に容易
- 検索窓を埋め込めたり、検索結果を埋め込めたりと、「検索」を自然にコンテンツとして埋め込むことが可能。

- データ構造に由来するものではないが、作ったキュレーションを他のWebサイトに張り付けることなども可能。
  →データをジャパンサーチに登録すれば、簡易的な検索サイトを簡単に作ることができる。

位置づけ

- キュレーションページは、SEOとして様々なメタデータ（引いては所蔵）への検索エンジンからの流入を増やすことに眼目がある。
- サイトに来た人がクリックだけで見ていくことができる（1クリック検索）
- パスファインダー的な意味合いも。
- 運用して半年では、効果は限定的。

- メタデータの個別ページも、SNS対応（OGP）等しており、Webの生態系上での流通をなるべく促進させたい。

4.その他の機能と今後の展望

画像系機能

Imgfeature

- 画像検索用の特徴量（1280次元のベクトル）を格納している。
- 画像検索→ID→Elastic searchとつなげることで、メタデータと組み合わせた絞り込みも可能。

Content

- 画像ファイルをアップロードすることが可能。
- 現状は、組織の画像や、キュレーションで使うための画像が主だが、原理的には
  各Itemの画像を載せることも可能。
- 画像はIIIFで配信。

その他

- 利活用という観点からきわめて有効なのは、ライセンスを明記していること。
- 他のデータベースとの横断検索や海外からの利用という観点では、RDFストアとSPARQLに対応していることが評価が高い。

今後の展望

- フィードバックやユーザログ分析からのUI見直し（機能多すぎ説）
- 検索や表示の高速化
- 将来的には、csearchやcurationのユーザ解放なども検討

- 今後は、使い勝手の向上とともに、連携先とギャラリーの充実等、コンテンツの拡充が大きな焦点（当館だけでできることではない）。
- 利活用事例の蓄積と、利活用方法に沿った機能開発をしていきたい。
- システムの機能が、少しでもインセンティブになれば。
  - XLSX+JPG→検索画面、API、正規化、RDF、IIIF、画像検索etc…

ご清聴ありがとうございました

ジャパンサーチ試験版（公式）
@jpsearch_go

フィードバックをお持ちしています