みんなで検索

一般財団法人人文情報学研究所　永崎 研宣

「みんなで翻刻」の活用例としての「みんなで翻刻サーチ」のご報告

- Web APIとしてのIIIFを活用し、さらに、みんなで翻刻Web APIを活用した事例
- ⇒さらにこの検索結果をAPIで他のサイトから使うことも可能

- 最近の比較的容易に導入可能な全文検索技術のご紹介
  - ⇒ちょっと使ってみましょうか。

システムの前提条件

- メンテナンスになるべく手間をかけたくない
  - できれば5年くらい放置しておきたい
- なるべくお金はかけたくない
  - フリーソフトで構築する
- なるべく手間をかけたくない
  - 構築運営手法の調査に時間をかけたくない
  - 構築作業に手間をかけたくない
- しかし、なるべく便利に使いたい

＝なるべく低コストな構築運用を

今回の検索に求められる要件

- 「みんなで翻刻」に含まれるテキストをなるべく便利に検索
- ⇒「うまく検索？」
  - ⇒何ができればうまく検索できたと言えるの？　？？？
- テキストの性格：
  - 著作権切れの近世くらい～大正時代くらいの⇒多様な時代
  - 古地震資料・その他日本史資料・仏教資料・草双紙・かるた…？⇒開放系
- ⇒とりあえず「何か発見できれば」よいのでは。
  - 全文テキストの横断検索による偶然的発見の機会を提供

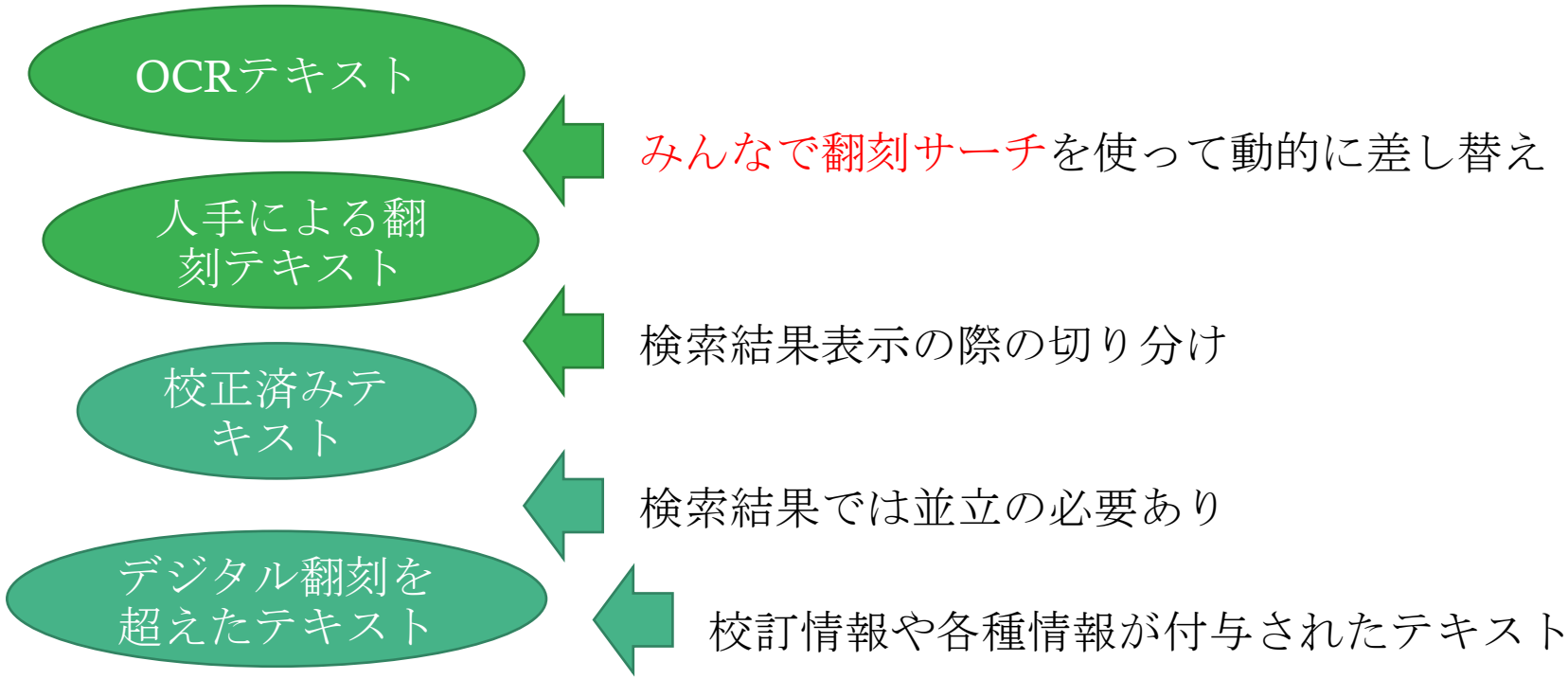
全文検索ソフトの選択

- 条件
  - 「システムの前提条件（2つ前のスライド）」を踏まえつつ…
  - 日本語で全文検索できるフリーソフト
  - 簡単に使えるもの
  - 横断検索による偶然的発見の機会を提供なるべくできるもの
    - 異体字同時検索／…？
- 検討したもの
  - Apache Lucene – Apache Solr
    - 枯れすぎではないか
  - Apache Lucene – Elasticsearch
    - バージョンアップがはやそう
  - Groonga
    - クロスプラットフォームが難しそう

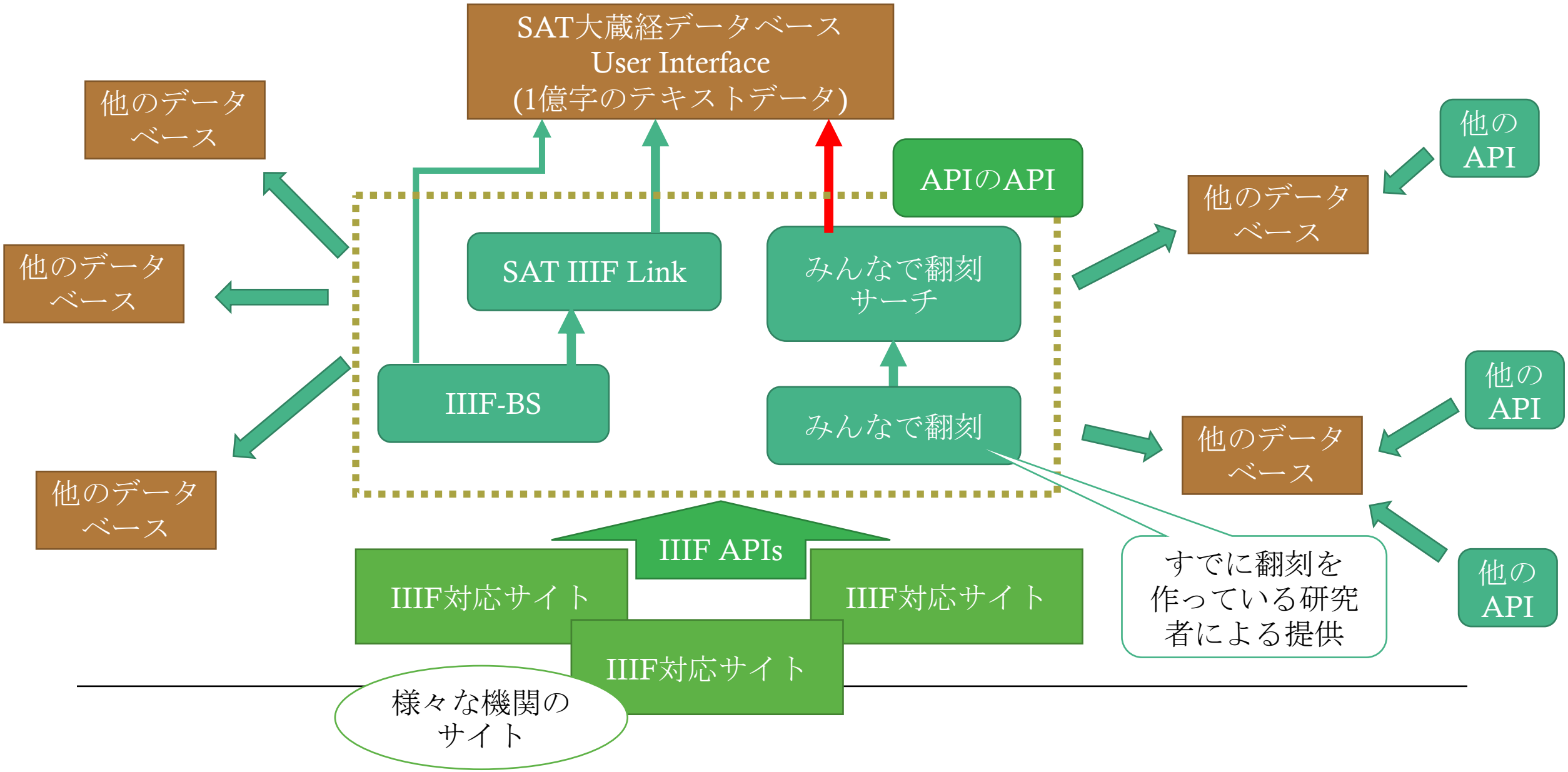
システムの構成

- 「みんなで翻刻」から自動的にテキスト取得
  - 現在は更新日付を見て更新があったものはテキストを取得して
    - ⇒橋本先生にAPIを作っていただきました
- Python3/FlaskでApache Solrを用いて検索
  - タグ等は全部削除して1つながりのテキストを作りn-gramインデックスで検索
    - 「形態素解析」をしない　⇒「東京」or「京都」と「東京都」が区別されずにヒット
      ⇒形態素解析の失敗に起因する取りこぼしは少ない
  - CHISE.orgの文字オントロジーのデータを用いた類似漢字同時検索
    - ⇒歴史史料に特化する場合、史料編纂所の異体字表の方が有効？
- インターフェイスには一部jQueryを使用
- 外部からAPI的に検索結果等のデータを取得可能
  - https://honkoku.dhii.jp/search/?url_key=%E4%BA%8C%E4%B9%98&url_var=1&url_proj=%E3%81%99%E3%81%B9%E3%81%A6&url_col=%E3%81%99%E3%81%B9%E3%81%A6&page=1&url_format=json

テキスト群の階層化


OCRテキスト
人手による翻刻テキスト
校正済みテキスト
デジタル翻刻を超えたテキスト
みんなで翻刻サーチを使って動的に差し替え
検索結果表示の際の切り分け
検索結果では並立の必要あり
校訂情報や各種情報が付与されたテキスト

デジタル仏典におけるAPIチェーンの一環として



どれくらい便利になった？

- 1時間おきに新規／変更データを本家から取得して検索
- 異体字で検索してもヒットする
- 「みんなで翻刻」全体でもコレクション単位でもアイテム単位でも検索可能
- Web API的な検索結果取得も可能
- 「前後数文字をまとめる」機能
  - ⇒「検索後のさらなる探索」へ
    - ⇒Apache Solrの検索結果を自分で細々プログラミングして処理
    - ⇒単語区切りが可能であれば共起情報がむしろ有用だが…
- Apache Solrの機能としては正規表現検索・複数語の近傍検索も可能だが実装に至っていない（検索インデックス作成に工夫が必要）
- 頁をまたぐ単語はまだ検索できない
  - ⇒検討中
    - 「みんなで翻刻」は単語／文章／段落区切りが明示的でな（くてよ）いのでやや難しい
    - ⇒入力者にお願いすると負担増