大学図書館による研究データ流通のサポート

(1)イントロダクション

(2)テキストマイニングXML論文の整理

2018.06.18 物質・材料研究機構 天野晃

(1)イントロダクション

ご注意

※実際の業務をベースに発表用にアレンジしています。
※未実装のものもあります。
※個人的提案及び見解として聞いてください。

(1)イントロダクション

弊職の業務のひとつに
- テキストマイニング用XML論文の整理

があり、具体的には
- データの取得
- 取得データの整理
  - 検索機能
  - アクセスビリティ
  - バックアップ

があります。
今日は検索を中心にお話しします。

※非公開

(2)テキストマイニングXML論文の整理

(2).1 データの取得とXMLの整理
- XMLデータの取得
- XMLタグ解析
- 全文検索

(2).2 XML以外のデータを検索可能にする(図・画像)
- 文字列(画像中の)
- Plot(Key:Value)
- 化学式(化合物)

※数式はMathML記述されており"ひとまず"処理不要

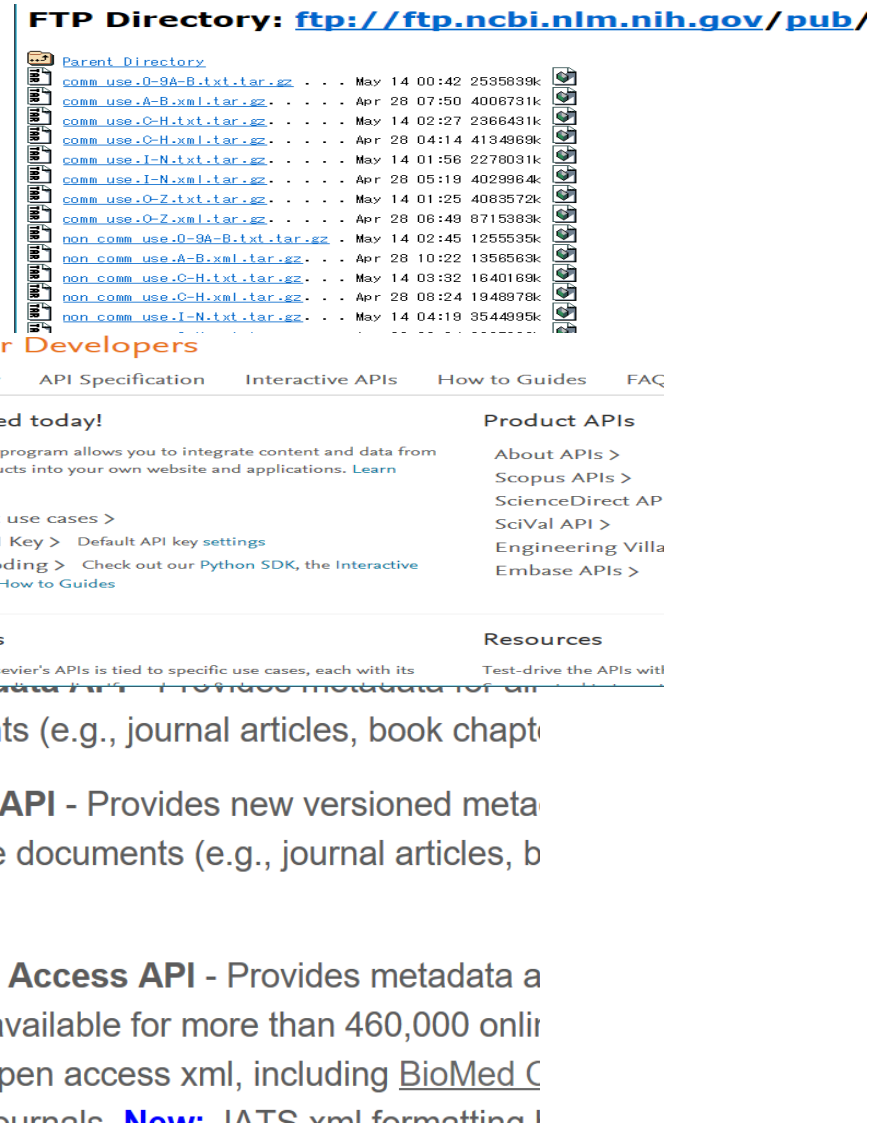
(2).1 データの取得とXMLの整理:XMLデータの取得

オープンデータ(論文)の例：

- NCBI PubmedCentral
  ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/
  ツール: 直接ダウンロード

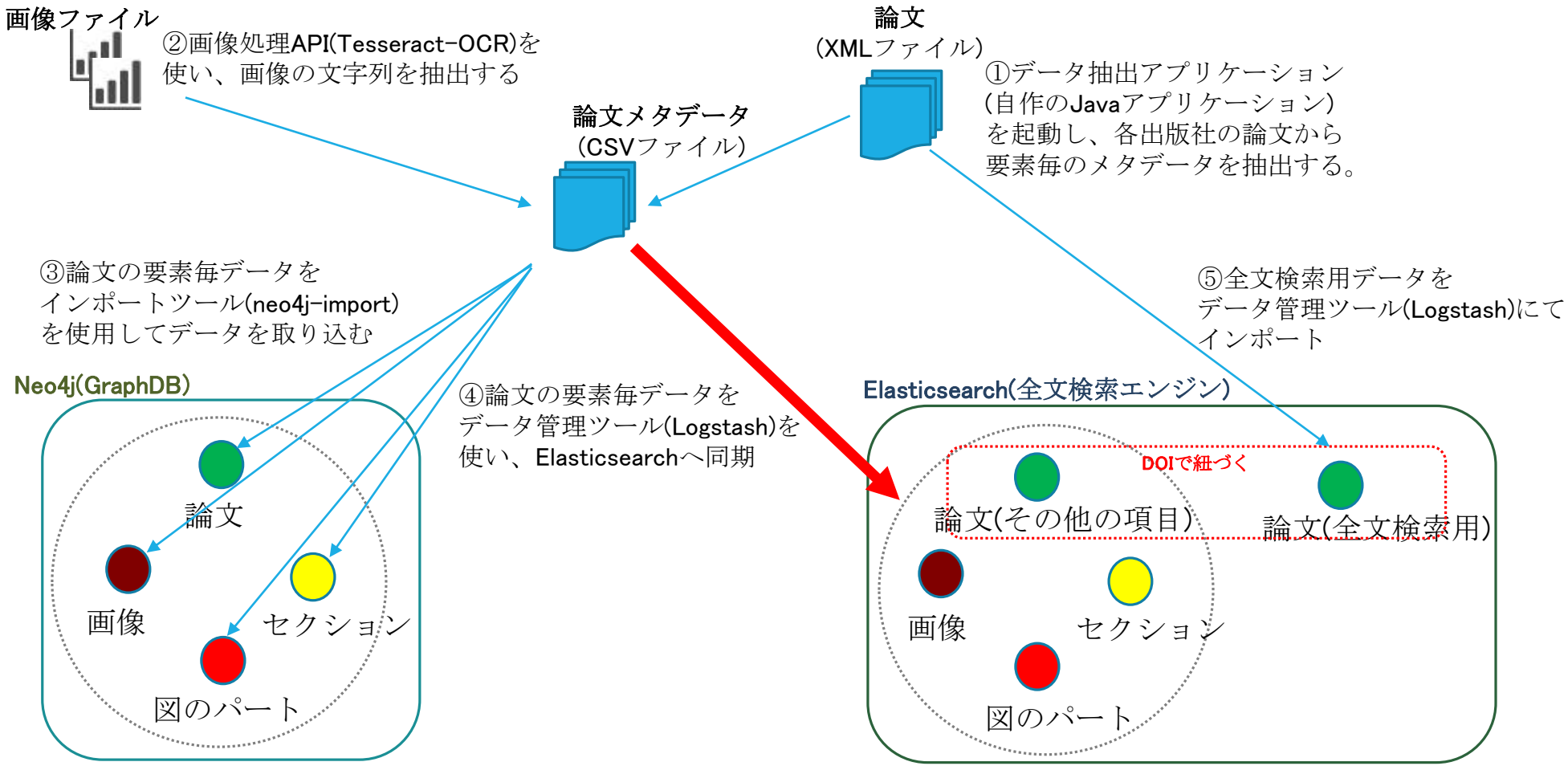
- Elsevier Developers
  https://dev.elsevier.com/index.html
  ツール: Python

- Springer Open Access API
  https://dev.springernature.com/
  ツール: curl + bash



※非公開

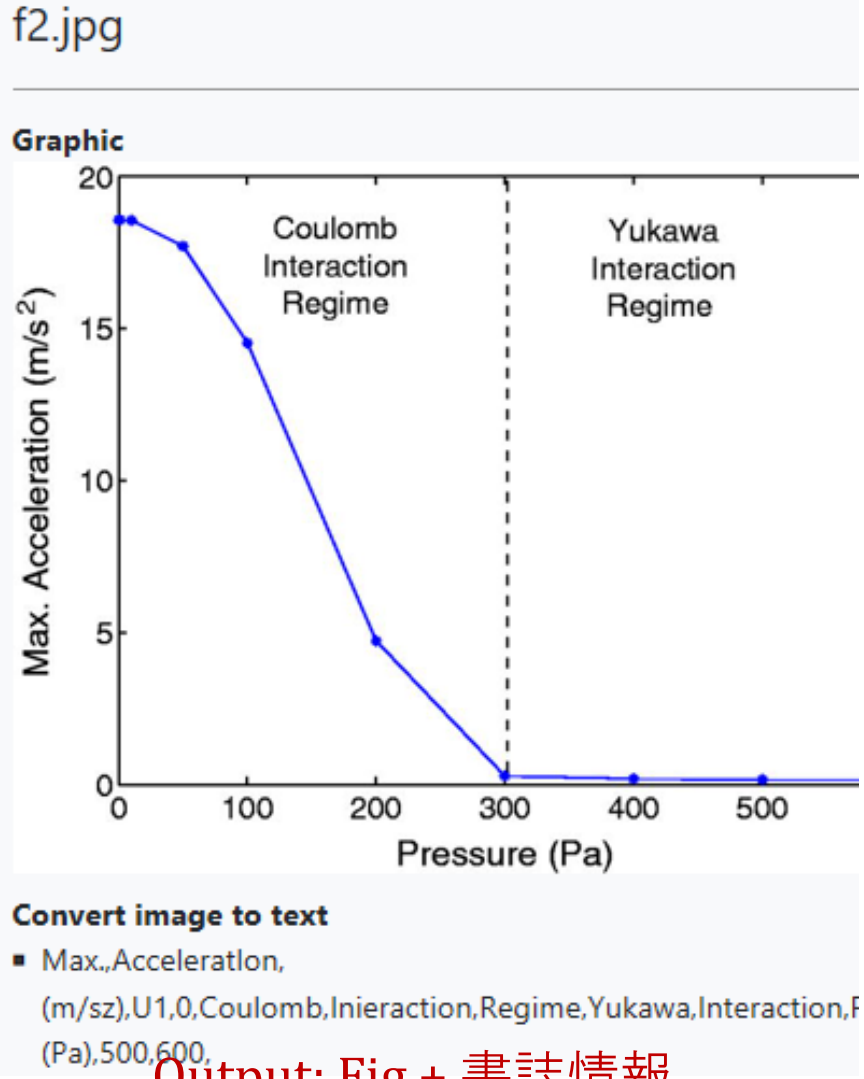
(2).1 データの取得とXMLの整理:全文検索



(2).2 XML以外のデータを検索可能にする :図中文字列

XML論文の図の文字列のインデクシング例:

1. XML論文と図のリンク処理(Logstash)
2. 図中の文字列の抽出(Tesseract-OCR)
3. 上記関係をインデクシング (Elasticsearch)



Yukawa → 全文検索システム

Input: Type=String

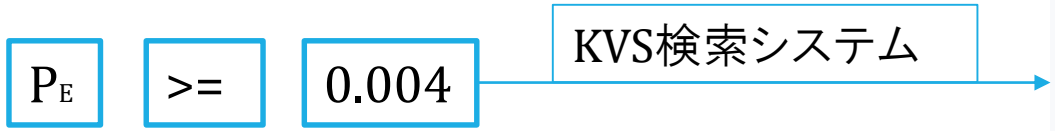
Output: Fig + 書誌情報

(2).2 XML以外のデータを検索可能にする:Plot

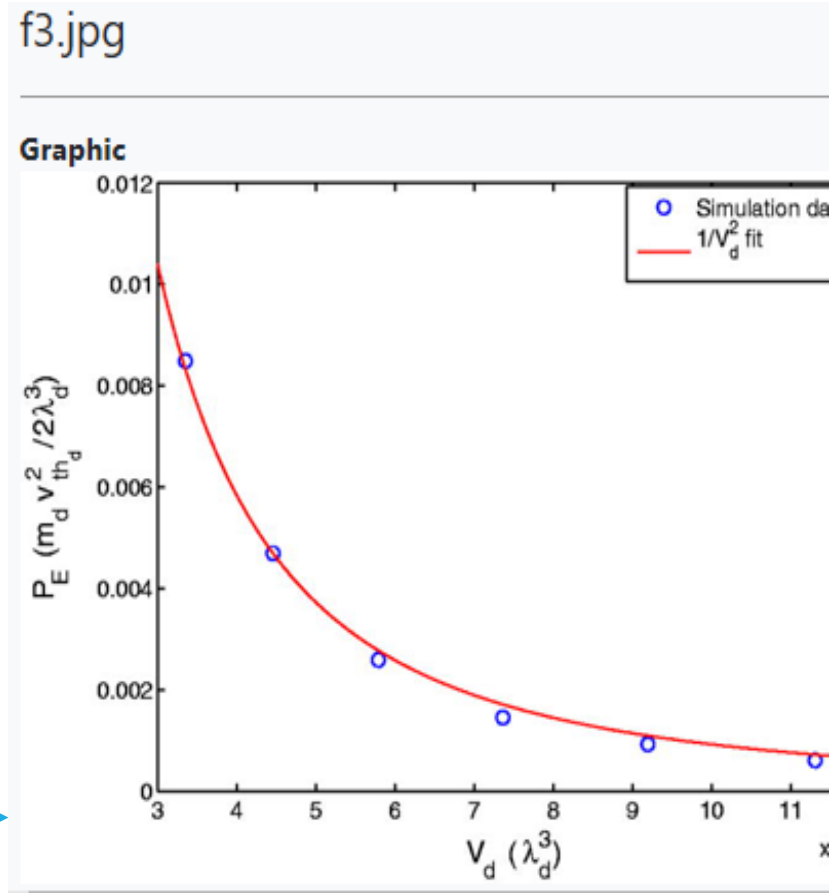
Plotからの[Key:Value]抽出例:

1. 線分と文字(数字)情報の分離
2. 線分情報から図内の複数plotを分離
3. 分離後各plotよりX-Y暫定値抽出
4. 数字の位置をヒントに暫定値を修正
5. DBへ投入(Elasticsearch)

※5以外すべて機械学習



Input: Type=Key:Relation:Value



Output: Fig + 書誌情報

(2).2 XML以外のデータを検索可能にする:化合物

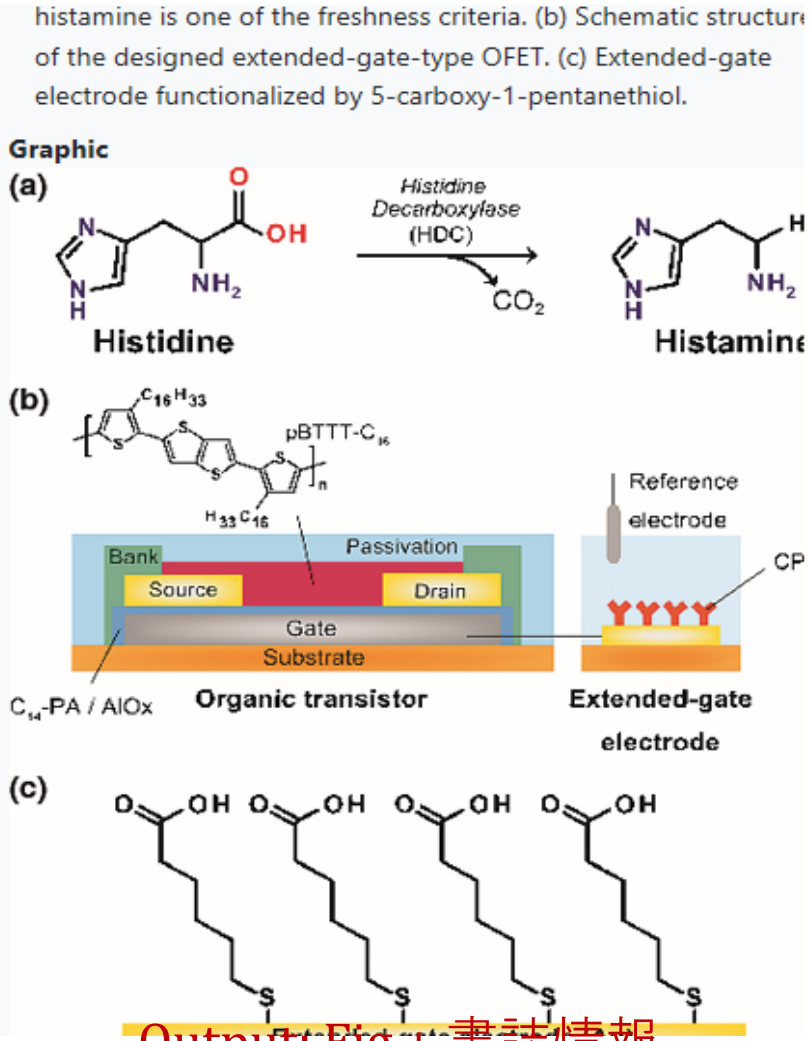
化合物検索例:

1.1 本文 => SMILES抽出(テキスト処理)
2.1 化合物表示の図を選択(機械学習)
2.2 図の化合物表示 => SMILES(OSRA)
3. SMILES => MOL変換(Open Babel)
4. MOL => 結合隣接行列(テキスト処理)
5. マッチング(次スライド)

C1=C(NC=N1)CC(C(=O)O)N

Input: Type=SMILES

化合物検索システム



Output: Fig + 書誌情報

※非公開