

(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

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。

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).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



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

FTP Directory: <ftp://ftp.ncbi.nlm.nih.gov/pub/>

File Name	Size	Modified
Parent Directory		
comm_use_0-9A-B.txt.tar.gz	2535839K	May 14 00:42
comm_use_A-B.xml.tar.gz	4006731K	Apr 28 07:50
comm_use_C-H.txt.tar.gz	2366431K	May 14 02:27
comm_use_C-H.xml.tar.gz	4134969K	Apr 28 04:14
comm_use_I-N.txt.tar.gz	2276031K	May 14 01:56
comm_use_I-N.xml.tar.gz	4029964K	Apr 28 05:19
comm_use_O-Z.txt.tar.gz	4088572K	May 14 01:25
comm_use_O-Z.xml.tar.gz	8715383K	Apr 28 06:49
non_comm_use_0-9A-B.txt.tar.gz	1255535K	May 14 02:45
non_comm_use_A-B.xml.tar.gz	1356563K	Apr 28 10:22
non_comm_use_C-H.txt.tar.gz	1640169K	May 14 03:32
non_comm_use_C-H.xml.tar.gz	1948978K	Apr 28 08:24
non_comm_use_I-N.txt.tar.gz	3544895K	May 14 04:19

Elsevier Developers

[My API key](#) [API Specification](#) [Interactive APIs](#) [How to Guides](#) [FAQ](#)

Get started today!

Elsevier's API program allows you to integrate content and data from Elsevier products into your own website and applications. [Learn more...](#)

1. Look at use cases >
2. Get API Key > [Default API key settings](#)
3. Start coding > [Check out our Python SDK, the Interactive APIs and the How to Guides](#)

Product APIs

- [About APIs >](#)
- [Scopus APIs >](#)
- [ScienceDirect AP >](#)
- [SciVal API >](#)
- [Engineering Villa >](#)
- [Embase APIs >](#)

Use cases

The use of Elsevier's APIs is tied to specific use cases, each with its

Resources

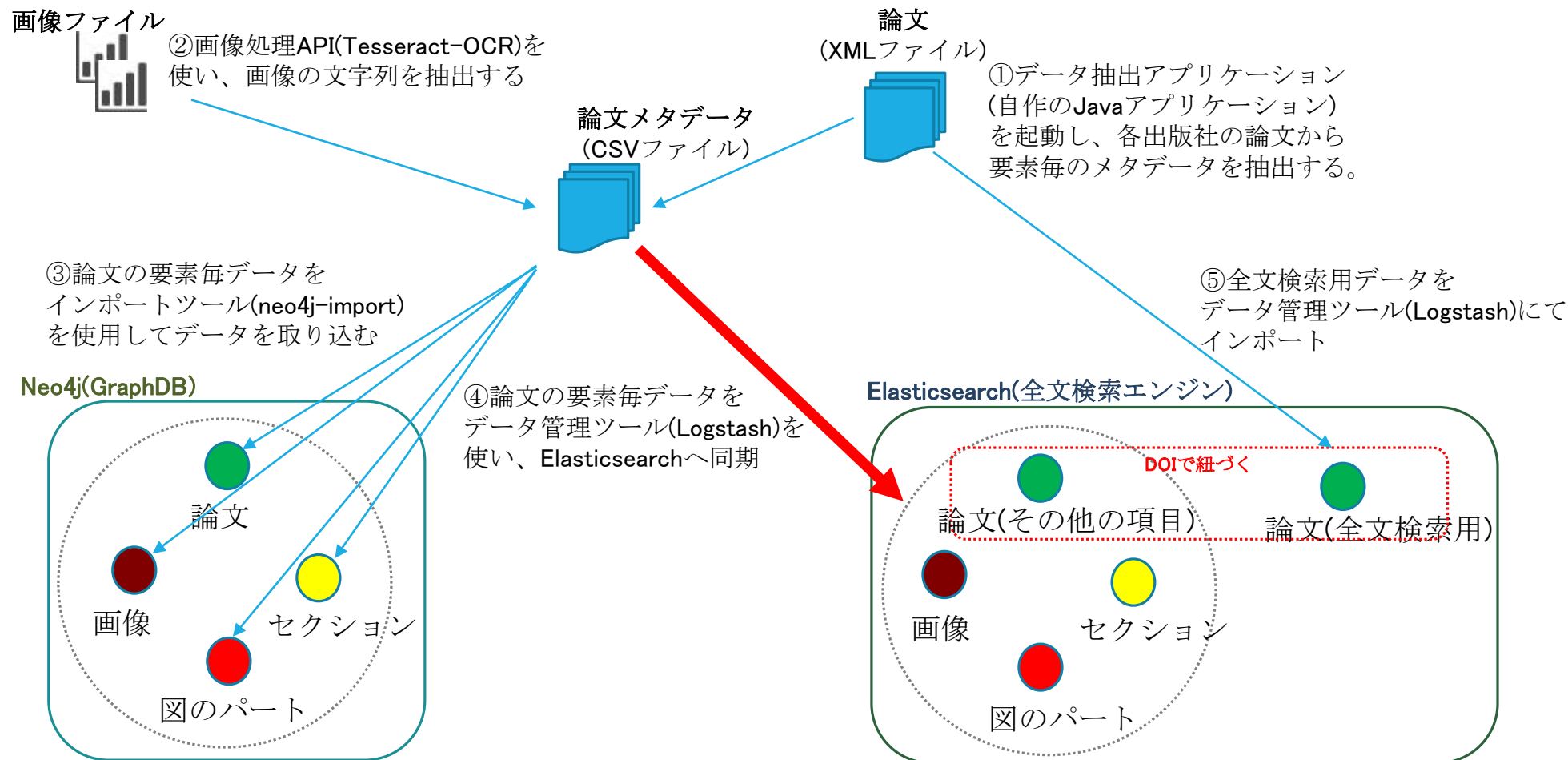
[Test-drive the APIs with](#)

Springer Meta API - Provides metadata for all online documents (e.g., journal articles, book chapters)

Springer Meta API - Provides new versioned metadata for 12 million online documents (e.g., journal articles, book chapters, protocols).

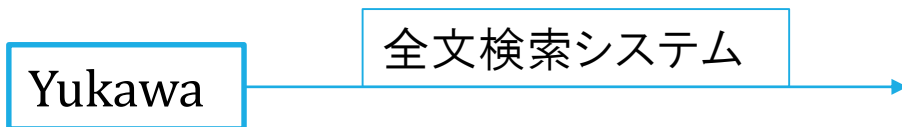
Springer Open Access API - Provides metadata for content where available for more than 460,000 online documents from Springer open access XML, including [BioMed Central](#) and [SpringerOpen journals](#). [New!](#) IATEX xml formatting!

※非公開

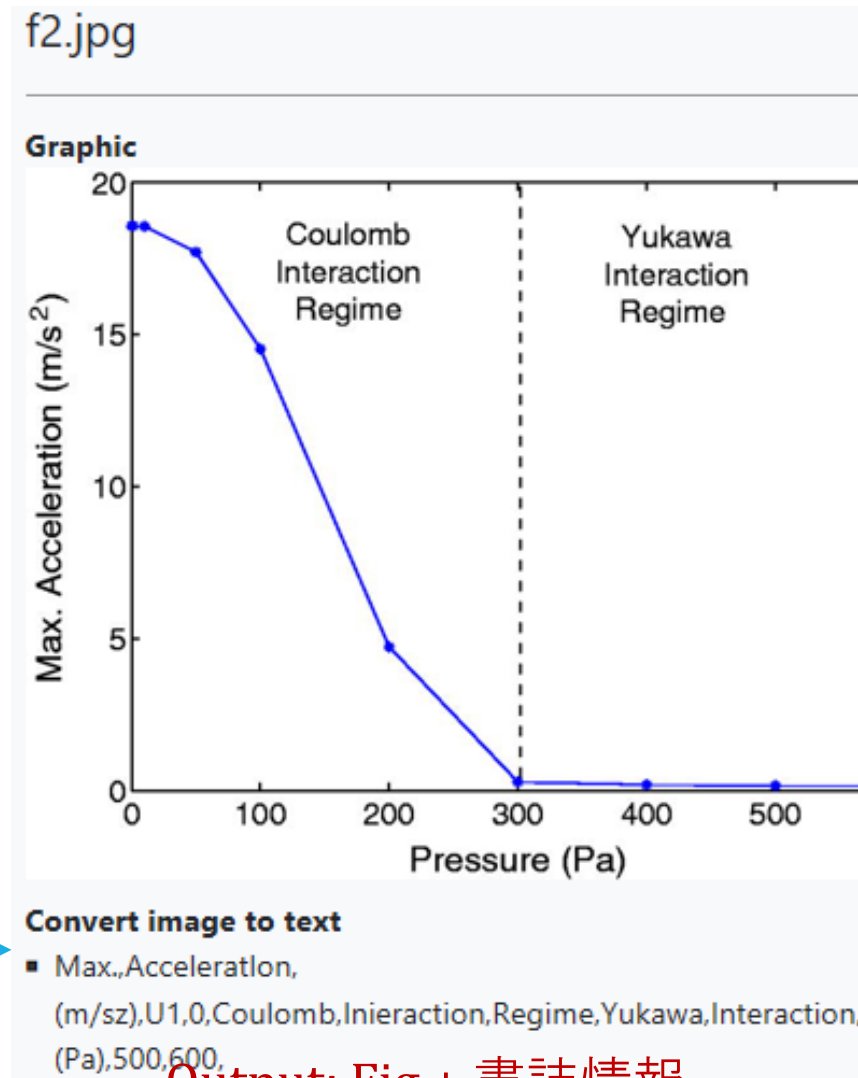


XML論文の図の文字列のインデクシング例:

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)



Input: Type=String



Output: Fig + 書誌情報

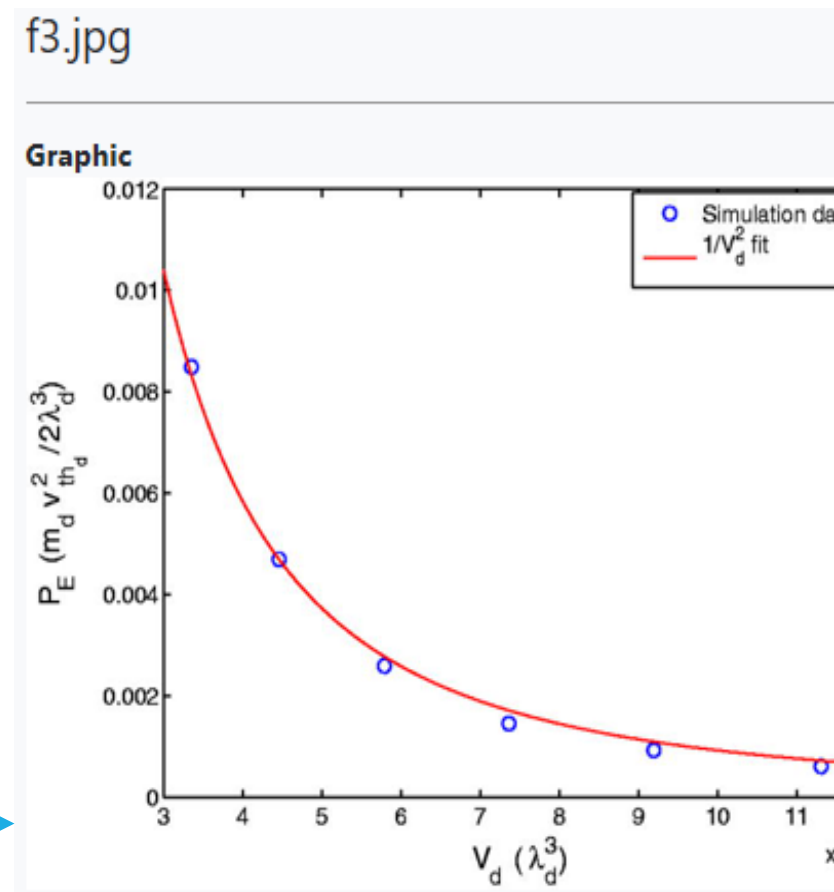
Plotからの[Key:Value]抽出例:

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

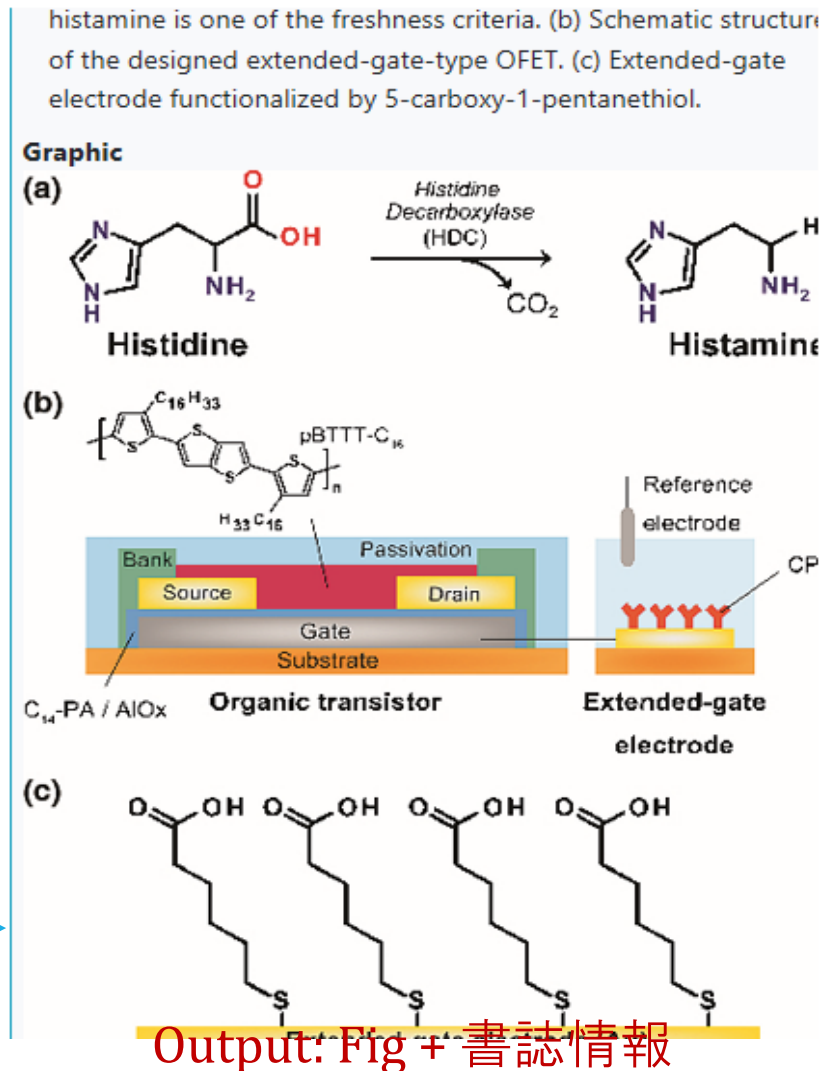
化合物検索例:

- 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



※非公開