

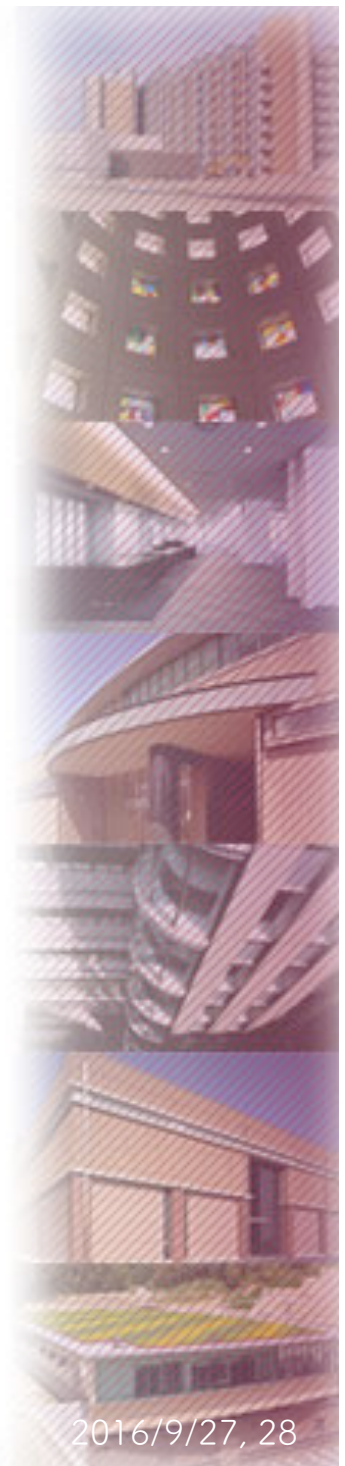
# 豊富な検索語で検索可能な データリポジトリの構築 に向けて

簡単なプロトタイプを構築  
その成果を報告

九州大学大学院 情報学専攻

池田 大輔

daisuke@inf.kyushu-u.ac.jp



## 動機：

異なる分野にわたる協同(地球課題解決型)

- 例：GEOSS (Global Earth Observation System of Systems)やFuture Earth
- リーディング大学院：決断科学

e-Scienceやオープンサイエンス

- 他分野の研究者によるデータの利用

シチズンサイエンス

- 一般の人による科学への貢献

様々な分野の人にも使えるデータリポジトリ

# チャレンジ

メタデータ  
による表現

機関リポジトリ：対象データ(論文)は抽象化が容易

- 図書館が扱ってきた図書と類似性が高い

データリポジトリ：対象データの抽象化が困難

- データは多種多様であり、汎用的な抽象化は厳しそう
- 一方、各分野で抽象化すると、他分野の人間には理解が難しい

それぞれの分野の「常識」がない人が利用

# プロジェクト紹介：FREEDxDOM

---

FREEDxDOM(フリーダム):

Foundation for REsEarch Data on Cross DOMains

科研費 基盤(B) 4年(2年目)

代表 池田(情報学)

分担 産学連携、天文、超高層大気分野の研究者

# プロジェクトのゴール：検索とクイックルック

## 検索

- 少ない前提知識で
  - 探索的に
- ⇔RDBは内部構造の知識を前提

## クイックルック：

- 即座のフィードバック

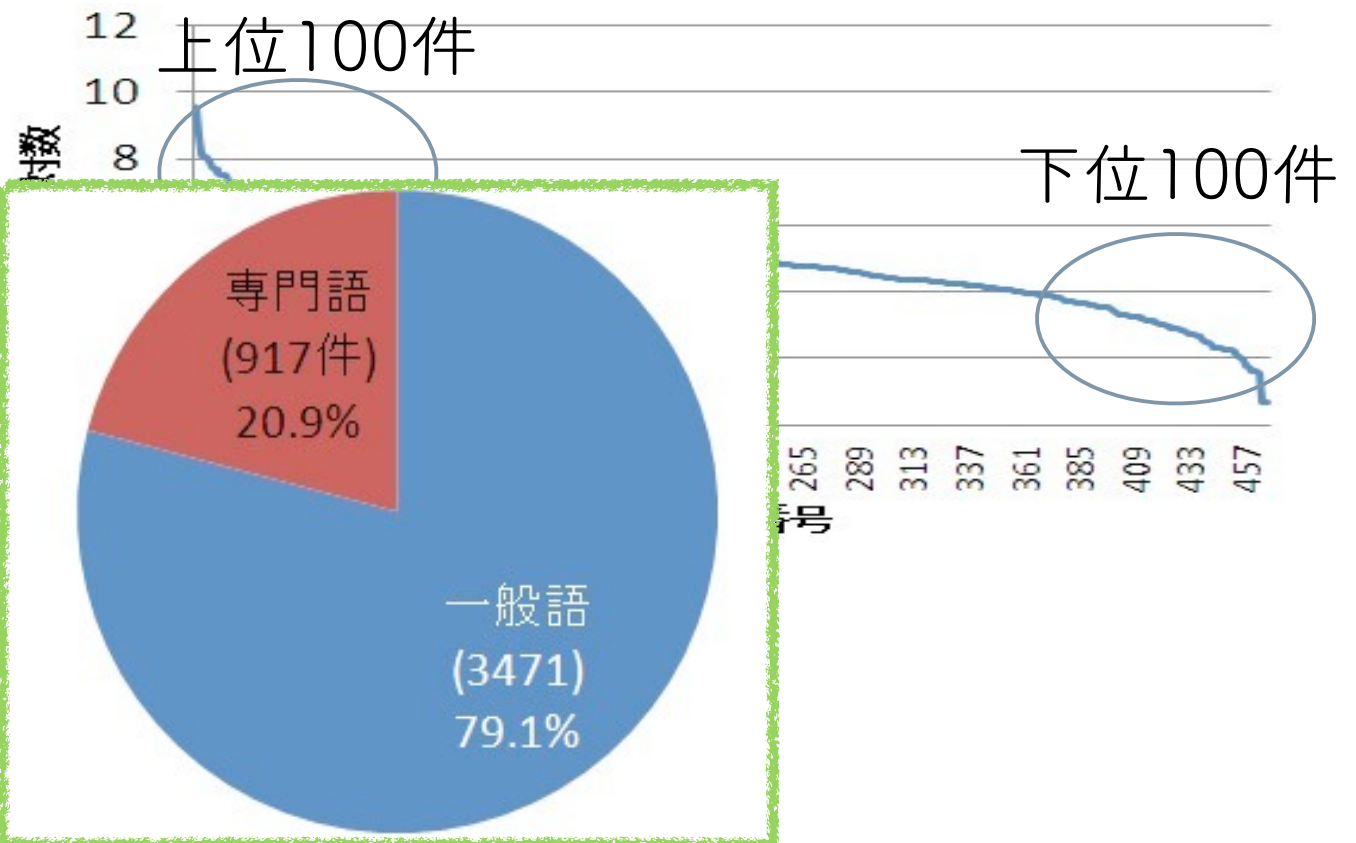


Google search results for "data repository". The search bar shows "data repository" and the results are filtered by "ウェブ". The search results show approximately 24,700,000 items found in 0.39 seconds. The first result is "Data repositories - Open Access Directory" from oad.simmons.edu. A red box highlights a snippet of text: "2015/09/01 - This is a list of repositories and databases for open data. Please annotate the entries to indicate the hosting organization, scope, licensing, and usage restrictions (if any). If a repository is open in some respects but not others, ...". Below this, there is a blue banner with the text "We can check the contents quickly." and another snippet of text: "Authors must deposit their data to a recommended data repository as part of the manuscript submission process; manuscripts will not otherwise be sent for review. We may recommend temporary deposition of your data to a general repository, ...". The second result is "WHO | The data repository" from www.who.int/gho/database/en/.

## (キーワード)検索とクイックルック

# アイデア：メタデータ vs. 検索

クエリ	ヒット数
ロールシャッハ FC:CF+C	43
日本の原子力発電 政策	7690000

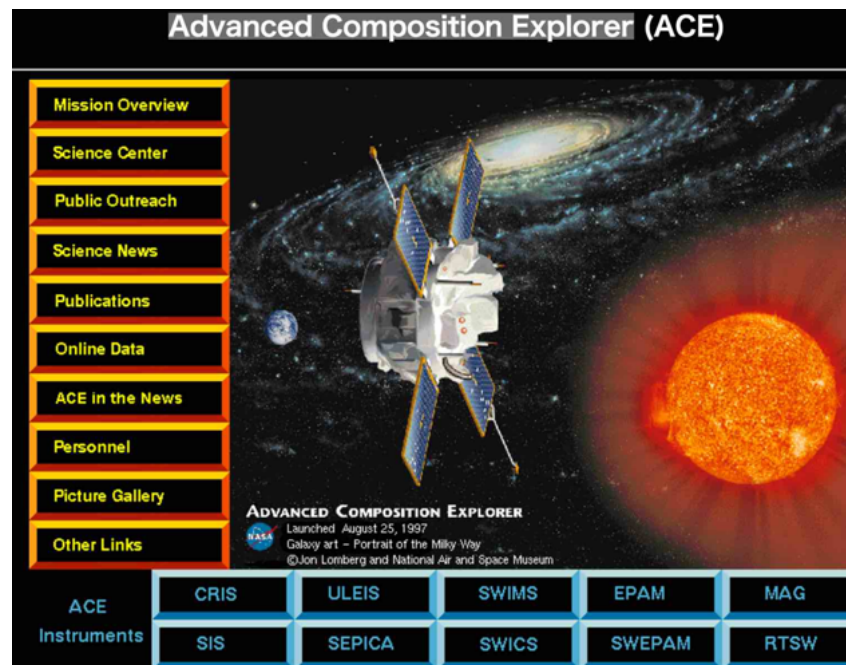


豊富な検索語(索引語)が重要

# 本発表の目的

特定のデータセットに限定して、一連のプロセスをまわす

- ACE (Advanced Composition Explorer)  
NASAの探査機で、取得したデータもACEと表記
- ACEに単語を紐づけ、データベース化



# プロセスの全体像

1. 地球物理分野の論文収集

2. ACEを使った論文の特定

注：「DOI付与と、これを用いた引用」は仮定していない

- まず、手作業でACEを使ったかどうか特定(ラベルづけ)
- 機械学習によりACEを使ったか論文かどうかを自動判定

3. 論文中の単語の重みを計算

4. 論文に単語のリストを対応づけ(DB化)

- 今回は、紐づけ対象のデータは一つなので、データではなく、論文に紐づけた

# 論文収集

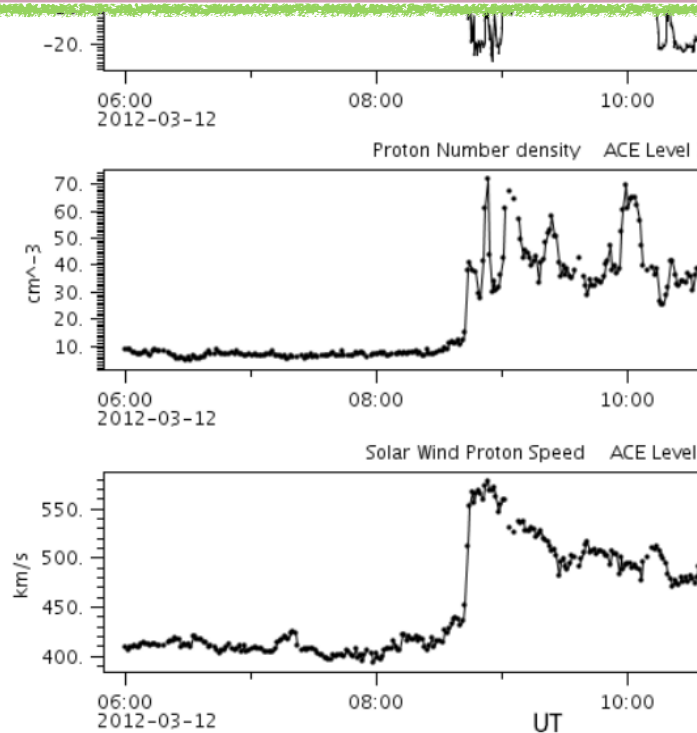
方針：ACEのサイト  
Publicationsから

- ACEを使ったか  
関係者の論文  
←使っていない  
場合も！
- 自動収集やマイ  
ニングの禁止の  
サイトが多い。
- 最終的には以下から  
Annales Geophysicae

オープンサイエンスデータ推進WS@京大

The image shows two overlapping website screenshots. The top screenshot is the 'Advanced Composition Explorer (ACE)' website, featuring a navigation menu with items like 'Mission Overview', 'Science Center', 'Public Outreach', 'Science News', 'Publications' (highlighted with a green dashed box), 'Online Data', 'ACE in the News', 'Personnel', 'Picture Gallery', and 'Other Links'. The background of the ACE site shows a satellite in space with the Earth and Sun visible. The bottom screenshot is the 'Annales Geophysicae' journal website, which is an open-access journal of the European Geosciences Union. It includes a search bar, a list of articles, and a 'Submit a manuscript' button. The URL <http://www.annales-geophysicae.net/> is displayed at the bottom of the page.

# ACEデータの使用・不使用の自動特定(1/3)



**Figure 2.** Solar wind variations during 12 March 2012 observed by the ACE space components, density, and speed are shown in three panels, respectively.

*Acknowledgements.* We thank the staff of EISCAT for operating the facilities. EISCAT is an international association supported by research organizations in China (CRIRP), Finland (SA), Japan (NIPR and STEL), Norway (NFR), Sweden (VR), and the United Kingdom (NERC). We thank the ACE SWEPAM instrument team and the ACE Science Center for providing the ACE data. Thanks also to the staff of WDC-C2 for providing the provisional geomagnetic indices. The solar image data obtained from the STEREO observations were provided by the STEREO Science Center/NASA. We thank the institutes who maintain the IMAGE magnetometer array and RSI/Rice University who provided the Boyle index. This work was supported in part by Grant-in-Aid for Scientific Research B (22403010, 23340144, 23340149, 25287126) by the Ministry of Education, Science, Sports and Culture, Japan. A part of this work was also supported by the joint research programs of the Solar-Terrestrial Environment Laboratory, Nagoya University and the National Institute of Polar Research, Japan.

Topical Editor K. Hosokawa thanks M. Kosch and one anonymous referee for their help in evaluating this paper.

データは図表で使用し、謝辞で言及

# ACEデータの使用・不使用の自動特定(2/3)

## Pythonの正規表現でACEに関する記述をパターンマッチ

```
# ACEの正式名称や提供元、含まれるデータセット名など
DataSetAbbr = re.compile("\\bACE\\b")
DataSetFull = re.compile("Advanced Composition Explorer")
DataSetProvider = re.compile("(ACE Science Center|CDAWeb)")
DataSetSubSets = re.compile("\\b(CRIS|ULEIS|中略|SWEPAM|RTSW)\\b")

# 構造(どこに現れたら「使用されている」とみなすか)
StructAck = re.compile("\\b(Acknowledgement|acknowledge)")
StructFig = re.compile("(Fig.|Figure)\\s+\\d")
StructTab = re.compile("(Tab.|Table)\\s+\\d")
```

## 機械学習の特徴(4×3=12次元)

- 上記の「構造」の後1,000文字にACE関連の記述  
=>1回とカウント

# ACEデータの使用・不使用の自動特定(3/3)

## 結果

	precision	recall	f1-score	support
0	0.70	0.88	0.78	8
1	0.95	0.88	0.91	24
avg / total	0.89	0.88	0.88	32

- precision(精度) : 出力した答えのうち、どれだけ正解か
- recall(再現率) : 正解のうち、どれだけ出力できたか
- F1 : 上記2つの調和平均

# まとめ

データへの索引語の付与の一連のプロセスを実行

- 論文収集
- ACEデータを利用した論文の自動特定
- 論文から索引語の抽出と重みづけ
- 索引語のDB化とインターフェイス



募集！

データと文献がセットになって、  
こんな検索エンジンが欲しいというところ

# 今後の課題

## 論文収集の問題点

- 多くの出版社サイトでは、自動収集や、収集した論文のマイニングを禁止

## データセットを表す語彙の学習

- 論文中のXYZというキーワードが何を表すのか？  
← 既存のメタデータベースのリポジトリが役に立つ??
  - DOI付きデータが引用されるようになると、このプロセスは不要になると思われる。

**プロがannotateしたフリーなリポジトリが重要!!**

# メタデータの有効性

## DML-JP：数学のナショナルポータル(北大 行木先生)

- NII支援による電子化
- IRのメタデータ
- コミュニティのメタデータ

The central window, 'DML-JP - Browse by Publication', displays a list of journals to browse from:

- Hiroshima Math. J. (2829)
- Journal of Mathematical Sciences, The University of Tokyo (271)
- Journal of the Faculty of Education, Kagoshima University (17)
- Journal of the Faculty of Science, Kagoshima University (105)
- Journal of the Faculty of Science, Shinshu University (6)
- Journal of the Faculty of Science, the University of Tokyo Sect 1 A (218)
- Kodai Math. J. (2034)
- Nagoya Math. J. (2232)
- Natur. Sci. Report. Ochanomizu. Univ. (205)
- Nihonkai Mathematical Journal (17)
- Publ. Res. Inst. Math. Sci. (2966)

Arrows indicate connections to:

- 機関リポジトリ (Institutional Repository):** Shown as the 'UT Repository' window, which is a digital archive of research outputs.
- Math. Reviews/MathSciNet:** Shown as the 'MathSciNet' window, which provides mathematical reviews and citations for research papers.

機関リポジトリ

Math. Reviews/MathSciNet

# 機械的な連携に有効

# 謝辞

今回の作業は、池田研の学生全員と、4日間の「池田研夏のプロジェクト」として実行しました。



学生さんの頑張りに感謝！！