

クライシス・ニュース・アーカイブをどう読むか？ 台風、東日本大震災、新型コロナウイルス感染症（COVID-19）の比較

ROIS-DS人文学オープンデータ共同利用センター（CODH）

国立情報学研究所

北本 朝展（Kitamoto Asanobu）

<https://researchmap.jp/kitamoto/>

@KitamotoAsanobu

クライシスとメディア

クライシスとは？

- ・ **クライシス（危機）** = 複雑なシステムが通常のように動作しなくなり、迅速な決断と対処が必要となっている状況。
- ・ **自然災害**（地震、津波、台風、洪水、地すべり、火山噴火、大雪）
- ・ **人為災害**（テロ、爆発、原子力災害）
- ・ **社会危機**（感染症、電力危機、経済危機、その他）

2012/07/19

国立情報学研究所 市民講座

8

クライシス情報学, NII市民講座, 2012年7月19日
<http://agora.ex.nii.ac.jp/~kitamoto/outreach/shimin-2012/>

1. クライシスの状況を刻々と伝える**膨大な情報**をどう**整理**するか？
2. 従来の**マスメディア**に加え、**ソーシャルメディア**が情報の多様性を拡大。
3. **アーカイブの役割**：日々の情報の流れから離れ、**現在から過去を俯瞰**する。

クライシス・ニュース・アーカイブ

<http://agora.ex.nii.ac.jp/crisis/news/>

1. Yahoo!ニュースに対して、固定キーワードで定期的に検索し、検索結果に表示される記事をクロール。
2. 「台風」：2003年5月～／11万件
3. 「地震」「原発」「東電」等：2011年3月～／65万件
4. 「肺炎」「コロナ」：2020年3月～（2019年12月まで可能な範囲で遡及）／64万件
5. 仕様変更にも可能な限り追従するが、網羅的な収集は困難。収集漏れは数パーセント程度？

ニュース・アーカイブの処理

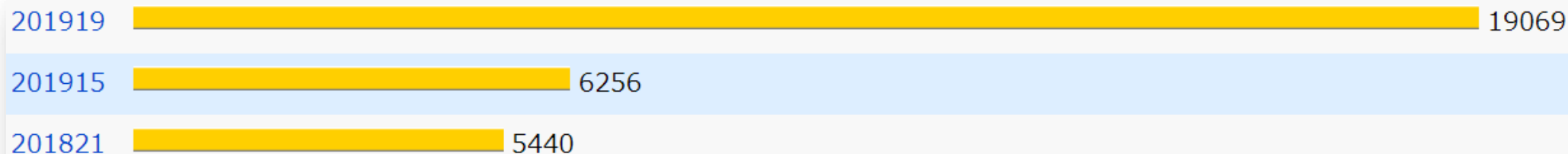
1. **前処理**：収集したHTMLからタイトル・日時・本文をスクレイピングし、テキストを正規化。
2. **形態素解析**：標準辞書に12万語以上の単語を追加。
3. **固有表現認識**：台風番号を認識、地名を認識（現在は運用停止中）。
4. **単語重要度評価**：時間スライスごとに単語の重要度を評価（TF-IDFに類似した尺度を利用）。
5. **ウェブサイト更新**：著作権の関係で本文は表示せず、分析結果の統計情報のみを提供。

台風ニュースと社会へのインパクト

<http://agora.ex.nii.ac.jp/topics/>

台風ごとの記事数は、社会へのインパクトの大きさを示す

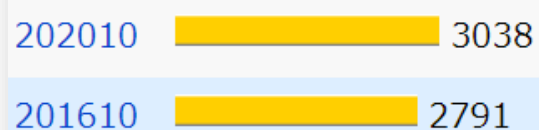
1. 令和元年東日本台風



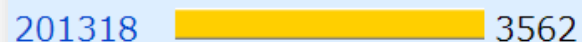
2. 令和元年房総半島台風



3. 関西地方強風



4. 紀伊半島大雨



東日本大震災アーカイブ

事数時系列グラフ



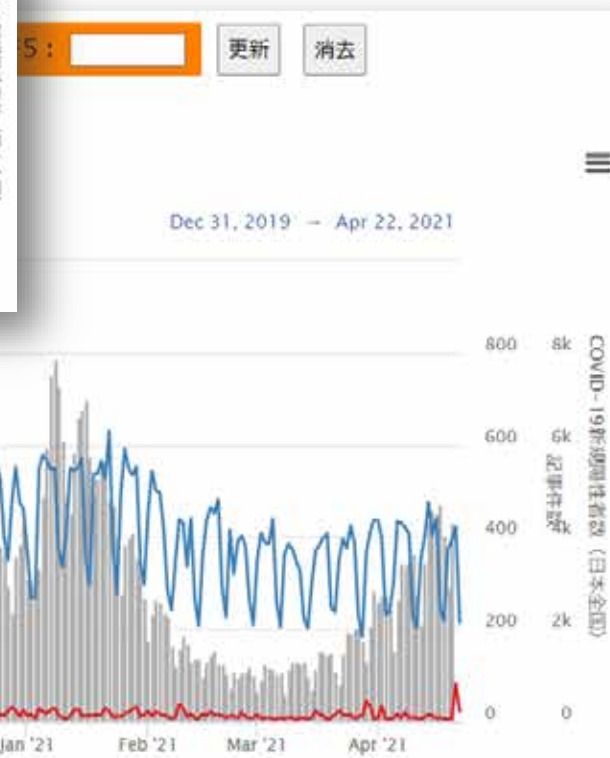
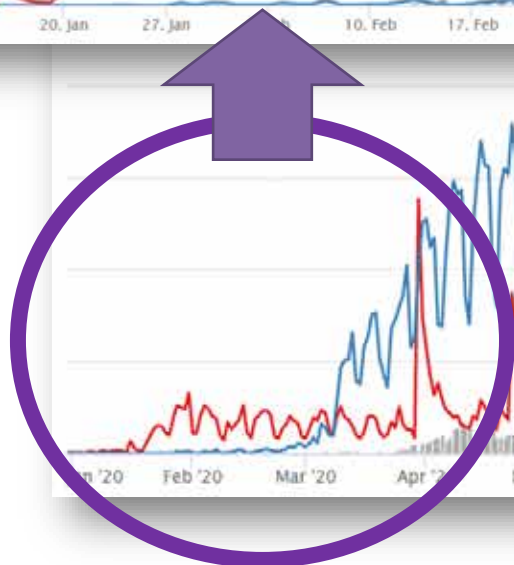
COVID-19アーカイブ

記事数時系列グラフ



(新型) 肺炎から「コロナ」への変化

赤：肺炎
青：コロナ



自然言語処理による分析

- **単語埋め込み（Word Embedding）**：テキストコーパス内で、**使われる文脈が類似する単語**を近い位置に埋め込むため、別表記や表記揺れ、概念的に近い単語なども取り出せる。
- 「東日本大震災」をキーとした場合：
 1. **地震コーパス**：他の震災が上位
 2. **台風コーパス**：他の風水害が上位
 3. **コロナコーパス**：災害のみならず社会変動なども上位
- コーパスの中心的な話題は**概念の解像度**が高く、周辺的な話題は概念の解像度が低いという傾向が見える。
- **コーパスを読む（遠読する）**ための自然言語処理ツールの一例。

東日本大震災タイムライン

1. 東日本大震災
2. 汚染
3. 野田
4. 予算
5. がれき
6. 再稼働
7. 値上げ
8. 原発
9. 衆院

<http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/mass-media/timeline/word/30-days.html.ja>

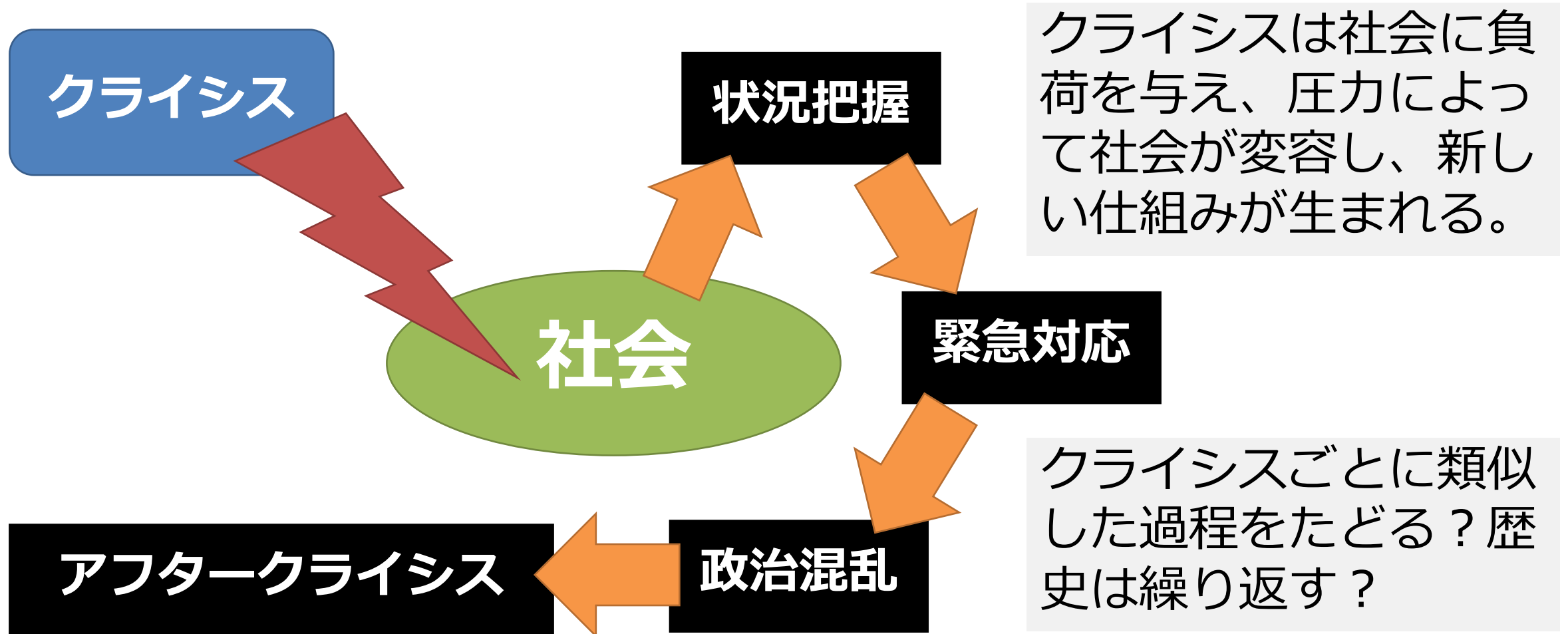
COVID-19タイムライン

1. 新型肺炎
2. 新型コロナ
3. 再開
4. コロナ
5. 菅
6. トランプ
7. バイデン
8. 感染
9. 接種

<http://agora.ex.nii.ac.jp/crisis/covid-19/mass-media/word/30-days.html.ja>

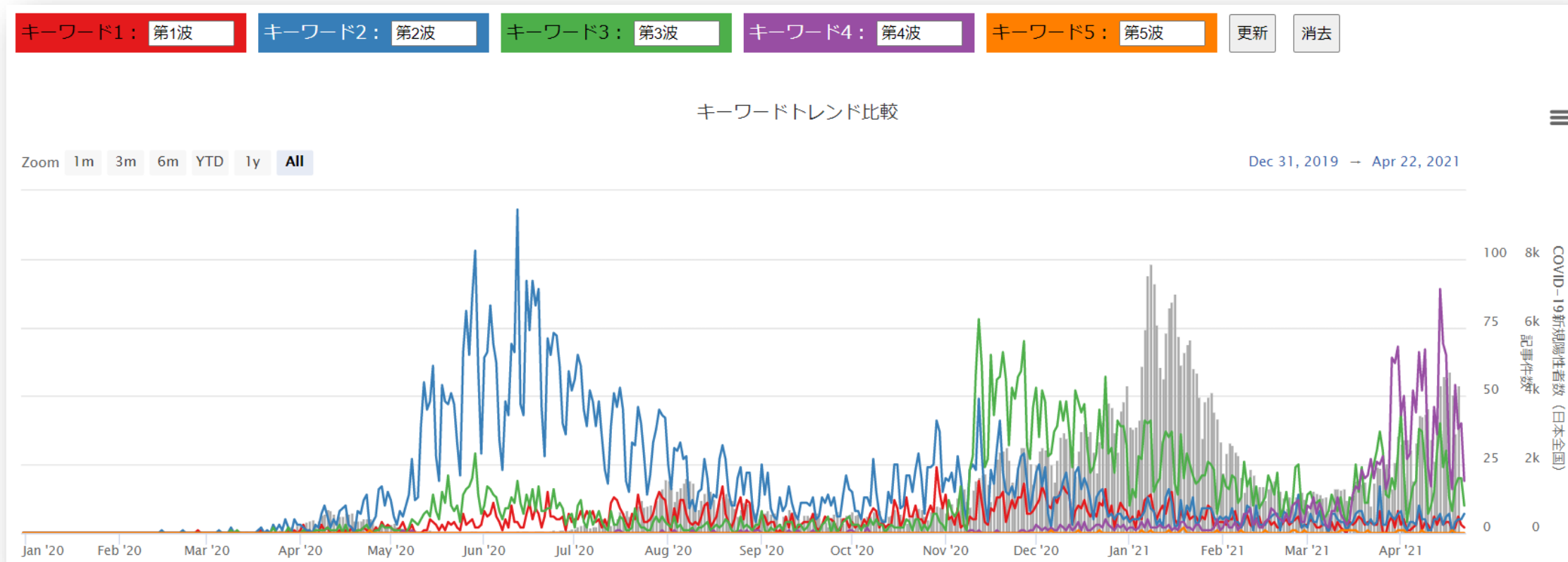
30日スライスの場合のトップキーワードの時間的变化。時間は上から下に流れる。「昨年」「来年」など、特定性が低い単語は除外。

クライシスと社会の変化



COVID-19 = ミニクライシスの続発

影響の局所性が強まらず、ニュース件数も高止まり



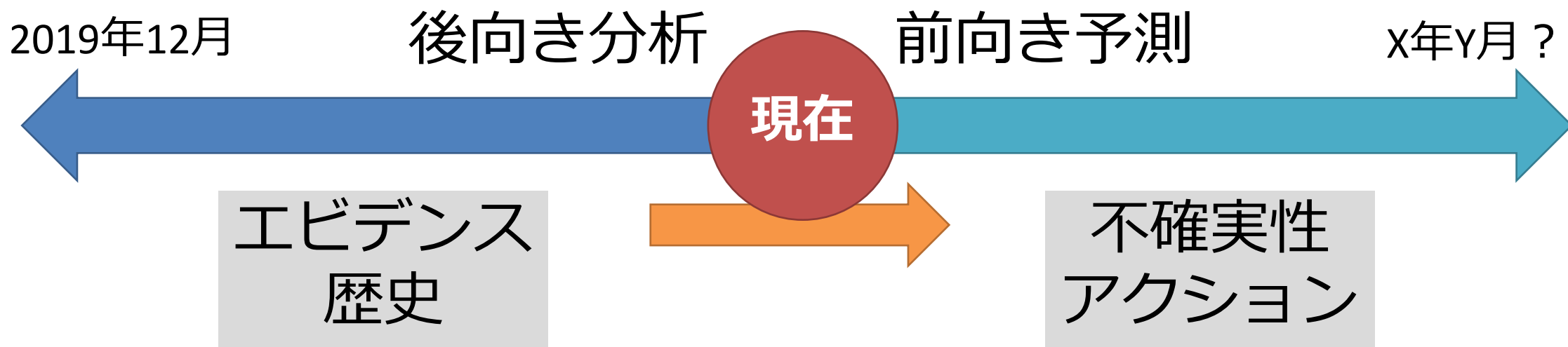
ニュース・アーカイブと歴史

ジャーナリズムは歴史の最初の草稿である
(Journalism is the first rough draft of history)

- **歴史（後向き分析）** = すでに出た結果を知ったうえで、整合性のあるストーリーを描き出す。
- **ニュース（前向き予測）** = これからの進展も想像しながら、現在の雑多な事象を記録していく。

新型コロナウイルス感染症（COVID-19）に関するニュース・アーカイブは、後世に書かれる歴史の素材となる

ニュース・アーカイブと未来



ニュース・アーカイブは、
後向き分析として書かれる
歴史（災害の教訓）の素材
という役割しかないのか？

ニュース・アーカイブから、
アクションナブルな情報を取り
出すために、どのような
整理・分析が可能か？

まとめ

1. マスメディアのニュース記事を、特定のクライシスに絞って収集したテキストコーパスを構築した。
2. 人間が詳細に読むには大規模すぎるコーパスを、機械の支援で理解しうる表現に圧縮することを試みた。
3. クライシスごとのタイムラインの比較、社会の変化の共通性、COVID-19の特異性などを分析した。
4. ニュース・アーカイブからどんなことを知りたいか？
ニーズから分析手法を考えていくことが重要。

関連ウェブサイト

- **新型コロナウイルス感染症（COVID-19）ニュース分析**
 - <http://agora.ex.nii.ac.jp/crisis/covid-19/mass-media/>
- **東日本大震災ニュース分析**
 - <http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/mass-media/>
- **デジタル台風：ニュース・トピックス**
 - <http://agora.ex.nii.ac.jp/digital-typhoon/topics/>
- **クライシス・ニュース・アーカイブ**
 - <http://agora.ex.nii.ac.jp/crisis/news/>
- 本研究の分析に用いたニュース記事は、Yahoo!ニュースから配信されたものです。