UDC2019 No. 188
小城藩日記プロジェクト

吉賀夏子

2020-01-31

アクティビティ部門

概要

小城藩日記プロジェクトは，**小城藩日記データベース**項目の一つである「記事」データを**「オープンデータ化」**するプロジェクト

小城藩日記データベースでは，江戸時代の**業務日誌の記事（タイトル）**を閲覧可能だが，現代の我々が読むには敷居が高い.

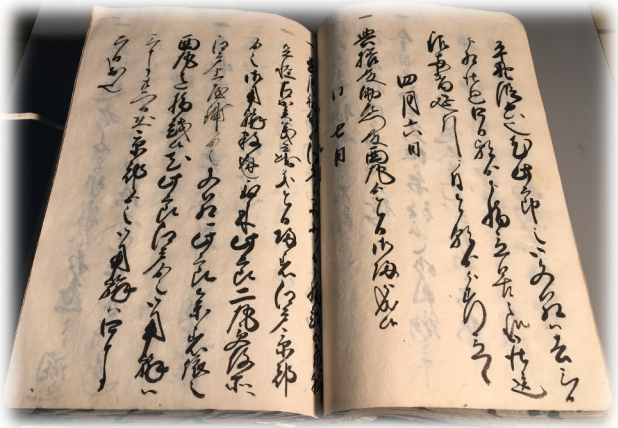
「小城藩日記データベース」とは

https://www.dl.saga-u.ac.jp/ogiNikki/

- **江戸期に作られた業務日誌のデジタルアーカイブ**
  - **（佐賀県の）小城藩**（佐賀藩の支藩）が記録した**業務日誌**「日記」のタイトル部分を検索できる
  - 小城藩は佐賀藩の支藩
  - 冠婚葬祭、佐賀藩との連絡や幕府、行政、経済、兵役、事件、災害など、当時の幅広い出来事を記録
  - 現在は専門家が「くずし字」から翻訳（翻刻）した**4万件以上のタイトル文と原文画像を閲覧**できる
  - 2019年11月に**丸善雄松堂「ゲスナー賞（知の組織化部門）」銀賞受賞**

- **オープンデータ**として2018年4月からウェブ公開中
  - **利用許諾：教育・研究利用は実質自由**
    - **原本画像を含む「全データ」を個人でダウンロードして、ホームページ，SNS，分析など気軽に利用することを想定**

佐賀大学附属図書館「日記」原本



小城藩日記データベースの特徴

- 異なる種類／組織のデータと組み合わせることができる機械可読データ**（リンクトデータ（Linked Data）形式）**版書誌データの積極的利用と発信を行う
  - 利用許諾を変更して**オープンデータにするよう附属図書館（所蔵者）と交渉**して実現
  - 異種データとの組み合わせで，**多様な人々の参画と多角的な分析**を期待
- タイトルの自動分類や年表／人間関係の可視化などデータの応用例を積極的に示す
- 和暦年月日をHutime（日時のLinked Datta）に接続し，ユリウス積算日およびグレゴリオ歴に変換することで，時系列データを構築
- 原本画像をブラウザでより快適に閲覧できる仕組みであるIIIF（トリプルアイエフ）をいち早く採用
- 今後も歴史学の研究者の協力でデータの拡充が進められる予定

オープンデータとは何か

- ``あらゆる人が自由に閲覧し、利用し、修正し、そして共有できることを **知識/*knowledge* がオープンである**とする。その際に掛けられる制限は、出自情報やオープンさの保持を考慮する程度に留められる。”

https://opendefinition.org/od/2.1/ja/

オープンデータと言えるデータにするには

あらゆる人，あるいは機械が…

利用可能な
ライセンス

アクセス
可能

データそのものが…

機械可読

ソフトウェアで
利用可能な
形式

https://opendefinition.org/od/2.1/ja/

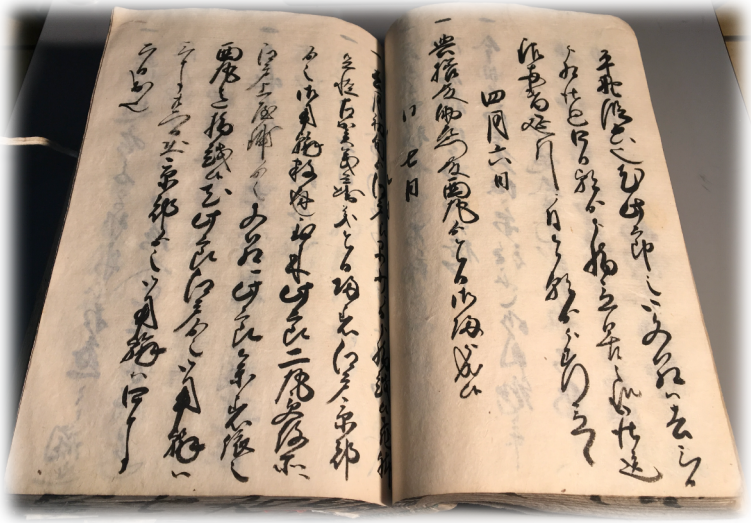
機械可読化

- Machine readable。コンピュータが自動的に読み込んで処理できるデータ形式のデータ。たとえばCSV,JSON, XMLなど。機械可読なデータは**構造化データ**でなければならない。比較参照：人間可読。

OPEN DATA HANDBOOK
https://opendatahandbook.org/glossary/ja/terms/machine-readable/

古文書のオープンデータ化の流れ

古文書（日記の記事文）



くずし字を読み取り
テキストに変換

ほんこく
翻刻

古文書読解ができる市民
による協力が必須

コンピュータで処理しやすい
データに変換

機械
可読化

IT技術が必須

翻刻例）延宝5年3月18日

多賀丸様御上京御供犬塚茂右衛門被仰付候事

（多賀丸様が上京されるので，お供に犬塚茂右衛門が命じられたとの事．）

機械可読化に立ちはだかる大きな課題

- **膨大かつ専門的なデータ**を**簡単に**機械可読データにする手法がない
  - 日本語（単語区切りがない，漢字かな混じり）
  - 江戸時代の書き言葉・文法の読み手が少ない
  - データそのものが，**ウェブであらゆる人が「利用」可能な許諾になっていないことがほとんど**
  - **特に地域色の強いデータの収集が困難**

ウェブの欠点あるある

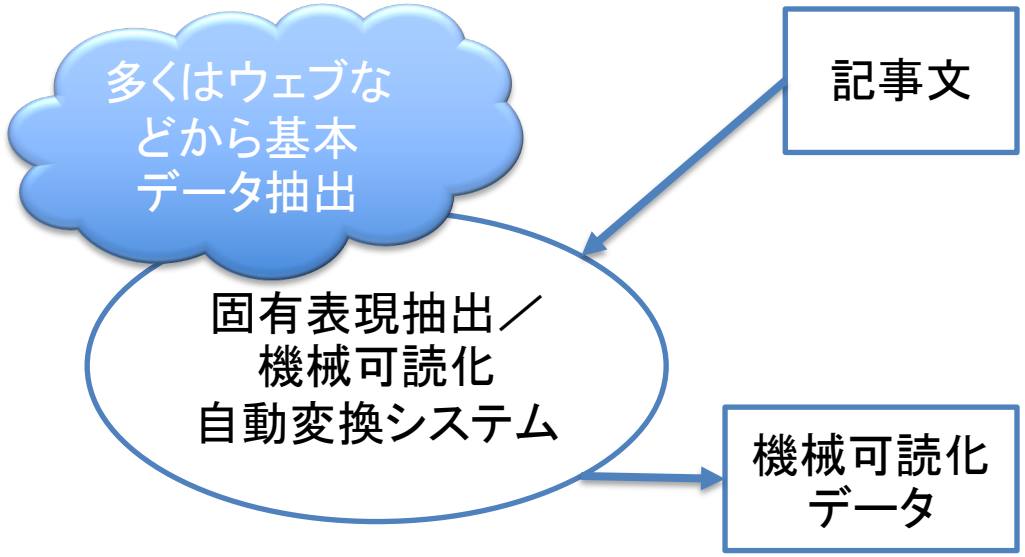
多くの人にとって比較的必要でない
地域固有のデータはなかなか見つからない

データを自由に利用していいか
分からないので**実質使えない**

課題解決のために：第1段階

- 小城藩日記データベースの利用許諾を**データ利用しやすい許諾に変更**するよう所蔵者（附属図書館）と交渉
- 記事文から**自動で固有表現抽出と機械可読化データへの変換を行うシステム**を作成
  - 固有表現を抽出するために、地域の人名、地名、役名、古文書の文法ルールなどの**基本データ**を収集
  - Githubに公開
    https://github.com/nikolito/ogiNikki

表示 - 非営利 - 継承 4.0 国際 (CC BY-NC-SA 4.0)

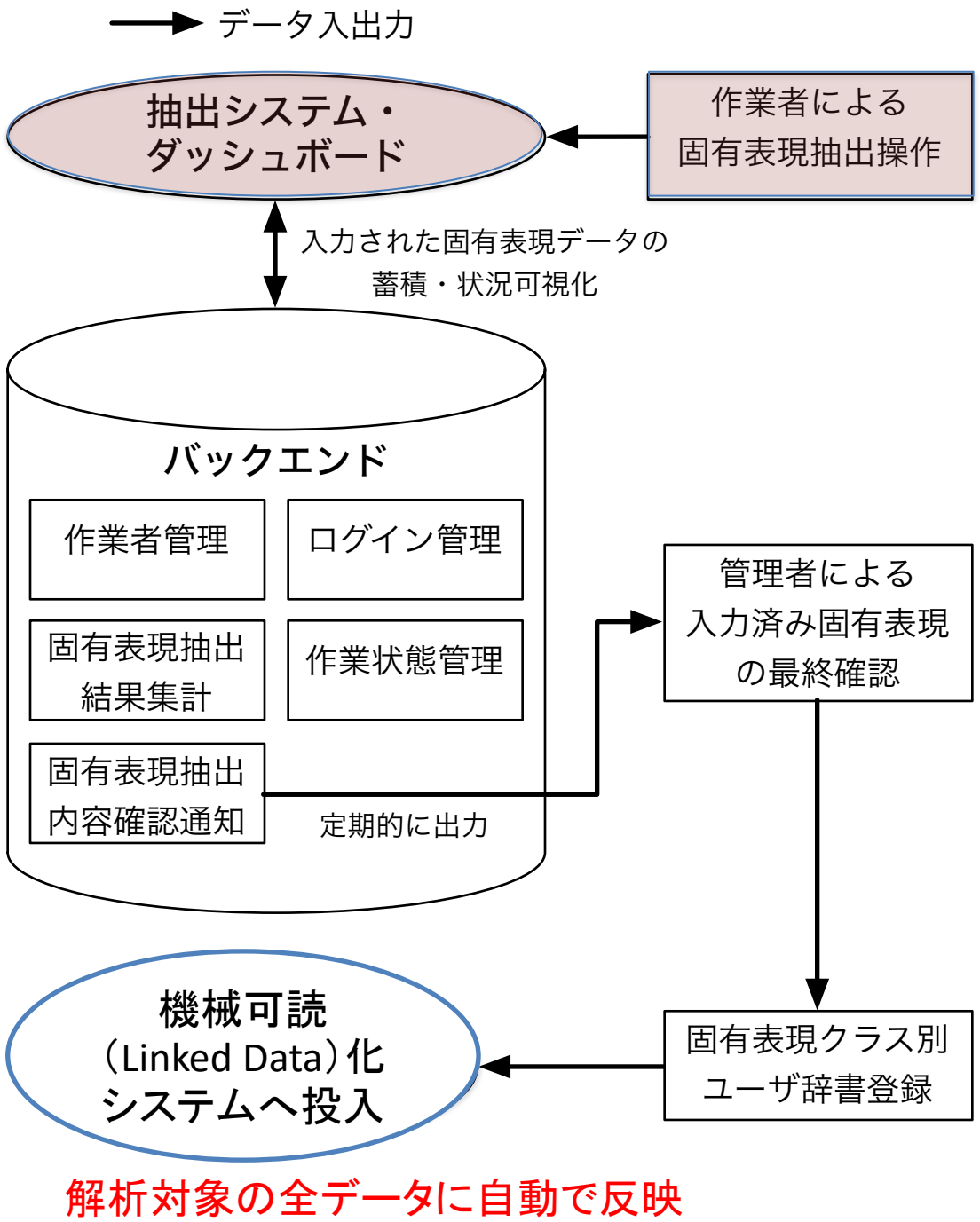


地域色の強い固有表現

キーワード

- 固有表現抽出
  - 資料中でキーワードとなる言葉を抽出して「クラス」に仕分ける
  - **人名・場所・出来事**は地域色が強く収集困難

固有表現クラス名		説明
Person	人名	人名，呼称
Date	日時	日時を表す語
Place	場所	座標で指定可能な地名
Event	出来事	検索キーワードとなり得る語
Role	役職、役割	役職，家族関係
Terms	候文用語	接続詞，定型句
Quantity	数量	数および単位を表す語



課題解決のために：第2段階

クラウドソーシングによる固有表現抽出システムを構築

固有表現が一つ作られると、データベース上で未探索の固有表現が全て抽出される

（固有表現）抽出システム作業画面例

ウェブブラウザで操作できるアプリを構築し実際に利用

小城藩日記データベース #24

翻刻文

多賀丸様御上京御供犬塚茂右衛門被仰付候事

単語切り出し結果

1. 多賀丸 名詞::固有名詞::人名::一般::*::*::*::*::多賀丸::*::多賀丸::*::固::*::*::*::*::OGI_JINMEI
2. 様 接尾辞::名詞的::一般::*::*::*::サマ::様::様::サマ::様::サマ::和::*::*::*::*
3. 御 接頭辞::*::*::*::*::*::オ::御::御::オ::御::オ::和::*::*::促添::基本形

ランダムに選ばれた記事文（翻刻文）および形態素解析（単語切り出し）結果の例

佐賀県立図書館データベース等の人名や地名データを使い，機械可読化はほとんど自動化済み

（応募者が構築）

翻刻例）延宝5年3月18日

多賀丸様御上京御供犬塚茂右衛門被仰付候事

人名　トピック（出来事）　役割　人名

文章を単語に分割するツール（形態素解析）による解析結果

1.多賀丸 名詞::固有名詞::人名::一般::*::*::*::*::多賀丸::*::多賀丸::*::固::*::*::*::*::OGI_JINMEI
2.様 接尾辞::名詞的::一般::*::*::*::サマ::様::様::サマ::様::サマ::和::*::*::*::*
3.御 接頭辞::*::*::*::*::*::才::御::御::才::御::才::和::*::*::促添::基本形
4.上京 名詞::普通名詞::一般::*::*::*::::::上京::::上京::::和::*::*::*::*::OGI_EVENT
5.御供 名詞::普通名詞::一般::*::*::*::::::御供::::御供::::和::*::*::*::*::OGI_ROLE
6.犬塚茂右衛門 名詞::固有名詞::人名::一般::*::*::*::*::犬塚茂右衛門::*::犬塚茂右衛門::*::固::*::*::*::*::OGI_JINMEI
7.被仰付候事 接尾辞::名詞的::一般::*::*::*::::被仰付候事::被仰付候事::::被仰付候事::::和::*::*::*::*::OGI_TERMS

抽出システム作業画面例2

あらかじめ固有表現を自動で抽出し、作業者に適宜修正を求める。

抽出システム作業画面例3

（判定に迷ったら作業状態を保存して再開・スキップできる。）

作業者は，固有表現抽出したい単語とそのクラスを「登録候補」にセットして、作業が終わったら「修正完了」ボタンを押す。

現在の小城藩日記プロジェクト

- 総勢8名
  - 全員**小城藩日記や古文書読解に一定の知識がある方**
    - 小城市立歴史資料館の古文書教室で紹介された方・職員
    - 佐賀大学地域学歴史文化研究センター職員
    - 私立大学講師
  - 4/8名は60-80代：簡単な説明と合同作業でフォロー
    - ウェブサイトの使い方はすぐに理解
    - **抽出ルールのすり合わせが重要**
  - 5/8名は有償で作業依頼

抽出ルールのすり合わせメモの例

固有表現ラベリング凡例（記事＃1～＃100 より抜粋作成）　R1.7.25.

①Person	②Place	③Event	④Role	⑤Terms	⑥Date	⑦Quantity	⑧
人(人名)	場(場所)	出(出来事)	役(役職役割)	候(候文)	日(日時)	数(数量回数単位)	固無(固有表現無し)

＃	記　事	固有表現分類
2)	長寿院様御死去之事　諫早石見殿内方直茂公御女	⇒ 人(長寿院,諫早石見),出(死去),役(内方，直茂公御女)
3)	城州様本行寺へ御葬礼之事	⇒ 人(城州),場(本行寺),出(葬礼)
4)	月堂様え御燈爐被相進候御方之事	⇒ 人(月堂),出(燈爐,御方),候(被相進候)　※「～被相進候御方～」「御方」だけ『出』に分類、「被相進候」は『候』へ
5)	月堂様御塔前へ御拝塔之衆之事	⇒ 人(月堂),出(塔前,拝塔)　※「之衆」は分類不要
6)	長寿院様御死骸諫早へ御越事	⇒ 人(長寿院),場(諫早),出(死骸),候(御越事)　※動詞入りの末尾単語は「事」まで入れ『候』へ→御越事，御書事，差迦候事
7)	南祥院様御塔前へ右同断事	⇒ 人(南祥院),出(塔前),候(右同断事)
8)	右御亡者様へ御名代諫早へ三浦杢之助御香典銀弐枚之事	⇒ 人(三浦杢之助),場(諫早),出(亡者,名代，香典，銀),数(弐枚)
9)	左近殿へ御同人御遺物鉄砲壱挺隼一連被遣候事	⇒ 人(左近),出(御同人,遺物,鉄砲,隼),候(被遣候事)，数(壱挺,一連)

※「御同人」の「御」外すと意味が違って来ると判断（目上の方の意）

「御」が目上の意で使われている場合は、「御」は Terms となります。

あらかじめ、抽出とクラス（この図ではラベリング）への
仕分け方について、作業グループで検討した。

固有表現リスト

https://winter.ai.is.saga-u.ac.jp/cs/ne-words

10 件/ページを表示

検索 肥州

#	固有表現種類名	単語（固有表現）	単語数
63	人名 Person	肥州様	399
3698	地名 Place	肥州	6
4830	人名 Person	肥州	4
6678	人名 Person	松平肥州	2
#	固有表現種類名	単語（固有表現）	単語数

1 ページ中 1 ページを表示 (全 11,521 件から 4 件抽出)

前 1 次

固有表現クラスへの分類作業情報共有

誰でも閲覧可能

作業者が、固有表現クラス判定に悩んだとき、参考にしている。

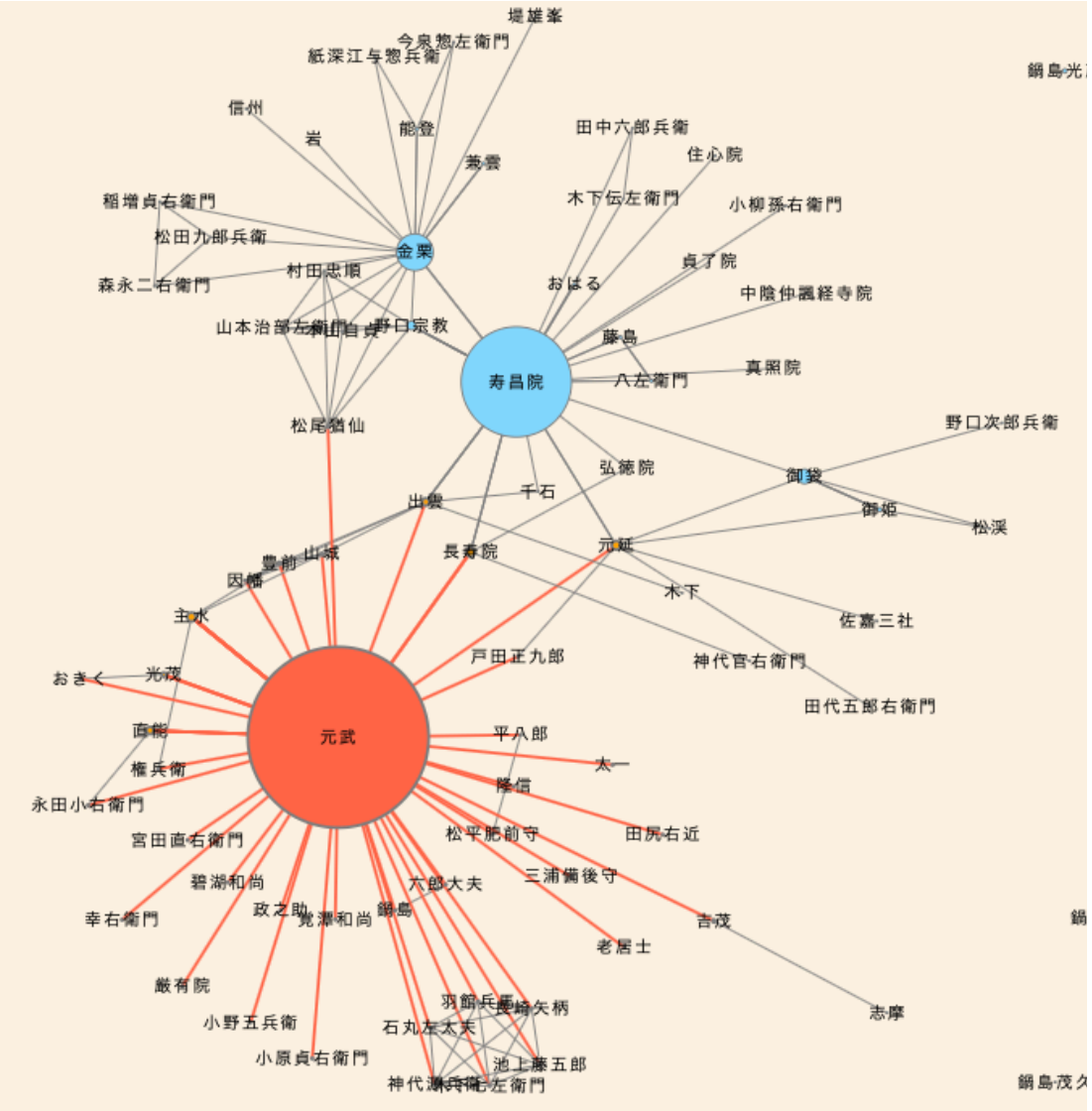
クラウドソーシングによる市民の協力と機械可読化システムの連携の相乗効果

- 佐賀県立図書館データベースなど、ウェブで一般的な固有表現の収集に加え、クラウドソーシングによる地域色の強い固有表現が追加されることで、**後から新しい記事文がデータベースに追加された場合でも**、現在の抽出精度と同等の固有表現抽出が自動で完了する。

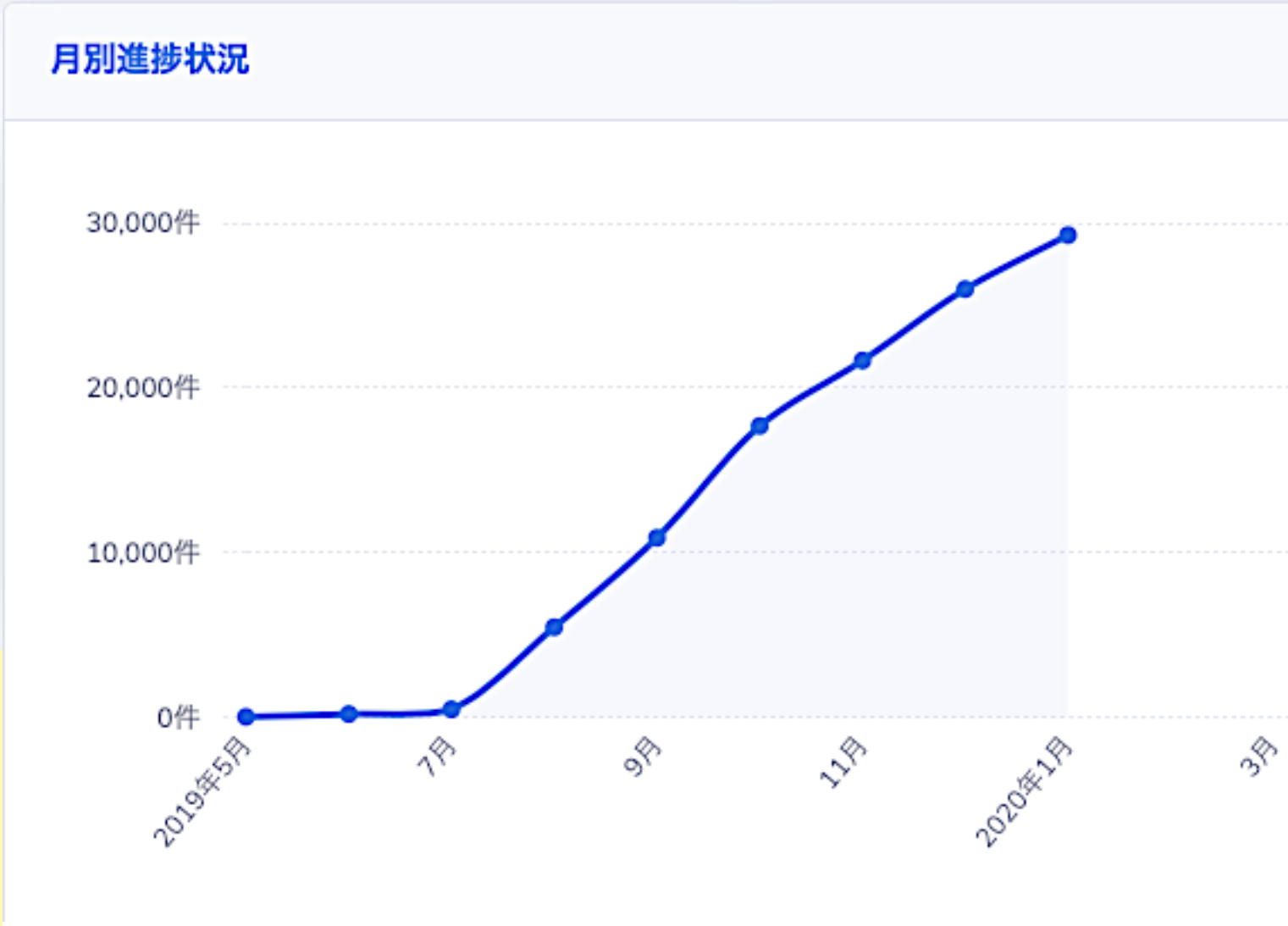
モチベーションの維持

- クラウドソーシングの性質
  - 簡易な操作
  - 短期間で興味を失いやすい
- 抽出した固有表現を，小城日記データベース記事文の**自動分類や年表，人間関係グラフ**などに利用
- 作業者各人が地元に馴染み深い人物間の関係あるいは地名，出来事などを検索し可視化することで，作業のモチベーションアップの一助としている．

（小城藩日記データベースでの）「元武」の検索で，日記上で同時に出現する人名関係を抽出／「元武」と同時に記事文に出てくる回数が多い人名ほど距離が短い→



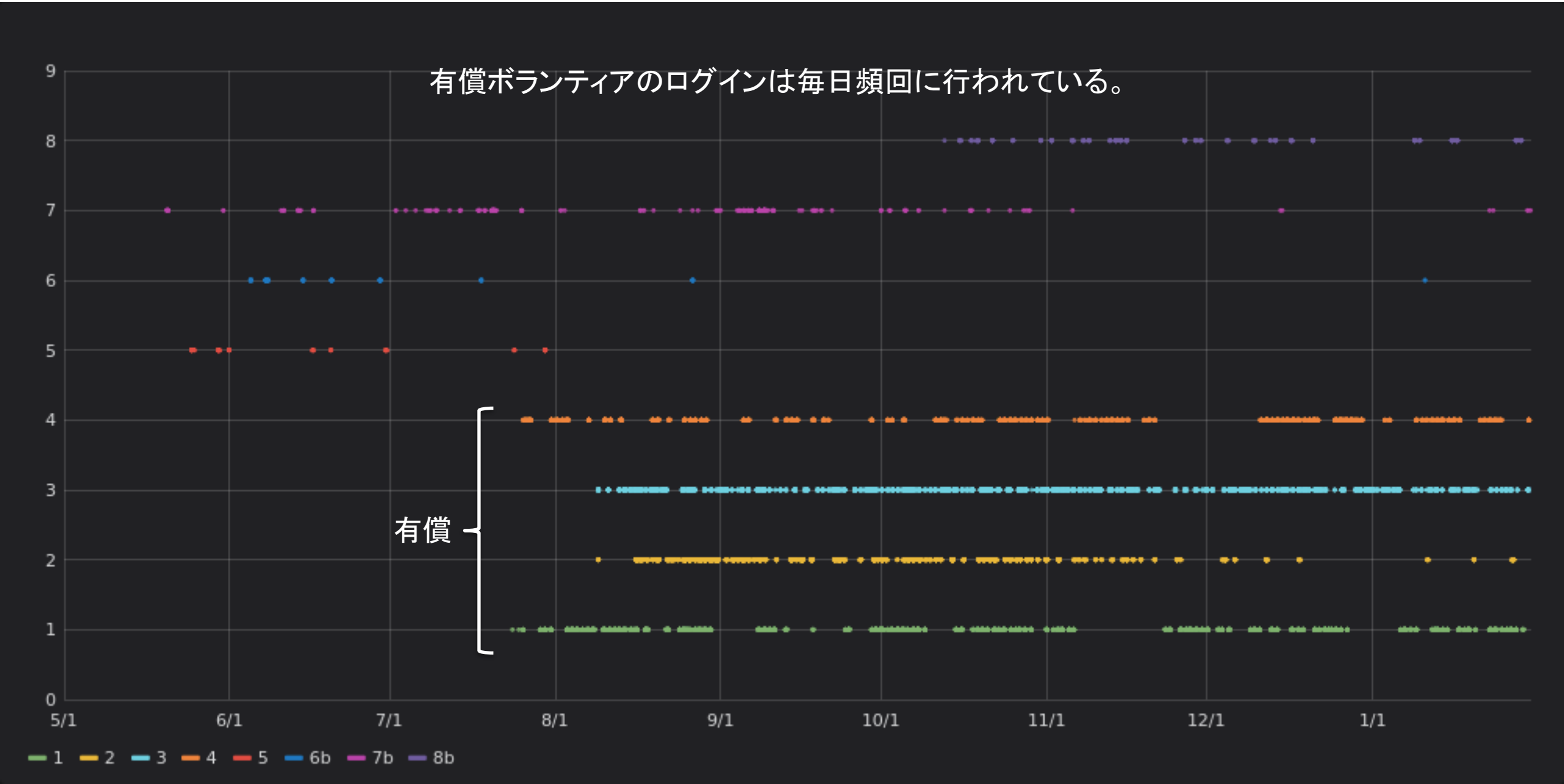
関係者全員がダッシュボードで成果を確認



2019年5月に3名で開始、7月以降に5名追加。専門性が高い作業にもかかわらず、8月以降に1日平均156件、これまでに記事文41719件中29265件の作業が当初の設定（30000件）を上回る速度で完了。そのため40000件まで作業する事になった。

固有表現抽出ウェブサイトへのログイン状況可視化

各数字は作業者につけたラベル・bが付く数字は無償、それ以外は有償ボランティアを表す。



一連のアクティビティまとめ

- 江戸時代の業務日誌「小城藩日記データベース」をオープンデータにした
  1. 附属図書館と交渉し，利用許諾をクリエイティブ・コモンズの仕組みで変更
  2. データを自動で機械可読化するシステムを構築
     - 翻刻された記事文から固有表現抽出し、機械可読化するプログラムを構築
  3. **地域色の強い固有表現抽出**を，郷土史の読み解きに熱心な地元**市民**を中心に依頼
     - 小城市歴史資料館，佐賀大学地域学歴史文化研究センターで，郷土史の読み解きに熱心な**市民8名**を「専門家」として募集
     - **クラウドソーシングによる固有表現抽出サイト**を構築し，10ヶ月稼働中
     - 固有表現抽出ルールのすり合わせを市民とともに検討
     - 全員の抽出結果はリアルタイムで共有し，作業の際の参考にできるよう公開

- クラウドソーシングによる市民の協力と機械可読化システムの連携で，後から新しい記事文がデータベースに追加された場合でも、全ての記事文に対して，常に現在の抽出精度と同等の固有表現抽出が自動で完了する。
- 全記事文から230984個の固有表現を抽出
- 作業に対するモチベーション維持として，関係者全員が具体的な作業成果をダッシュボードで確認可能
- 作業を有償ボランティアにすると，ほぼ毎日ログインがあった.
- 人間関係や自動年表，自動分類など作業成果を使った可視化もモチベーション維持のひとつになっている.
- 専門性が高い作業にもかかわらず、8月以降に1日平均156件、これまでに記事文41719件中29265件の作業が当初の設定（30000件）を上回る速度で完了。そのため40000件まで追加作業する事になった。