低コストな Linked Data 化を目指したクラウドソーシングによる固有表現収集の試み

吉賀　夏子・只木　進一（佐賀大学 理工学部）

歴史資料の分析には，多くの人手が必要である．人々の協力を得るための手段の一つとして，クラウドソーシングの枠組みに，学術研究の分野においても関心が寄せられている．本研究では江戸期の郷土資料から，クラウドソーシングで Linked Data 構築に必要な固有表現を抽出する手法について報告する．日頃から郷土資料に関心のある協力者は，使いやすい GUI を用いて，提案された固有表現およびそのクラスの候補を修正する．その結果， Linked Data 生成の過程を短時間で実行できるようになった．

A Low-cost Named Entity Extraction from Japanese Historical Documents in Edo Era through a Crowdsourcing Framework for Generating Linked Data

Natsuko Yoshiga / Shin-ichi Tadaki (Faculty of Science and Engineering, Saga University)

Analyzing historical documents requires a large amount of human efforts. In this context, crowdsourcing frameworks have been attracting interests even in academic researches for gathering cooperates from public. This paper reports a crowdsourcing framework for extracting named entities from Japanese historical documents of a local area in Edo era. Cooperators, who have been interested in local historical documents, correct suggested named entities through a convenient GUI. The framework has accelerated the process of generating Linked Data.

１．はじめに

様々な分野において，データの活用が進み，非数値データに関する取り組みも行われつつある．例えば，国立国会図書館や人文学オープンデータ共同利用センターなどが，文化財に関する所蔵資料のオープンデータ化を進めている[1][2]．こうした取り込みに対応して，人文学においても，歴史的地震記録の研究など，資料のデジタル化を超えて，デジタルデータを活用した研究が広がりつつある[3].

資料のデータ化だけでは，データの意味や構造をデータ作成者以外が知ることはできない．データは，その意味や構造を伴って，初めて知識となり，活用の可能性が拓かれる．

Web 上で特定のデータセットに関する知識を共有する一般的手法として，メタデータを Linked Data 化する手法がある[4]．表形式のデータに較べ，作成にやや手間のかかる Linked Data 形式の利点は，テキストであることと、データセットの中にデータ構造とメタデータ間の関係を内包することができることである．このため，Linked Data 形式のデータは，アプリケーション依存が無く，長期保存や共有に適している．

これまで，筆者は，江戸時代の古典籍コレクションに対して，書誌情報の注記からメタデータの Linked Data 化を行なってきた[5]．その際に，固有表現を事前にユーザ辞書に登録し，形態素解析を行うことが有効であることを示した．一方，ユーザ辞書への登録作業は，候補の選定から固有表現クラスへの分類まで人手の作業に依存している点が課題となっている．

佐賀大学地域学歴史文化研究センターでは，佐賀藩の支藩である小城藩の業務日誌「小城藩日記目録」の翻刻を行なっている[6]．筆者は，この事業と並行して，これらのメタデータ化を小城藩日記データベース上で行なっている[7]．「小城藩日記」は業務日誌であることから，小城周辺の地名，人名，行事など，地域色の強い固有表現を多く含み，一般的な用語辞典などが有効に活用できない．そこで，地域色の強い固有表現の抽出，分類，登録を，地域市民の協力の下で行う仕組みを提案する．

提案手法は，一般的な固有表現の抽出，分類，登録に係る作業を簡素化・半自動化することによって実現する．そのため，単に市民の協力を得る目的に留まらず，Linked Data 化の効率化に寄与するものである．

２．本研究の目的

本研究では，筆者が提案する自然文を含めた書誌情報に対して低コストに Linked Data 化を行うシステム（図 1，以下 Linked Data 化システムと呼

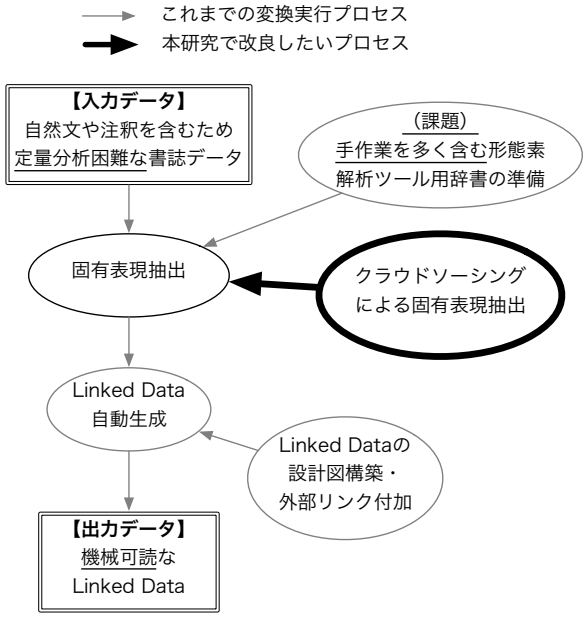

図 1　筆者が提案する自然文を含めた書誌情報に対する Linked Data 化システム[5]でのクラウドソーシングによる固有表現抽出の位置付け.

ぶ.）の一部である固有表現抽出の過程を改良する.

始めに，Linked Data の有用性を，書誌データを例として，表形式と比較して述べる．表形式の書誌データは，データ整理は簡便である一方，書誌項目の意味および用法まで明示的にデータ化されている例は少ない．特に，古典籍や古文書のような文化財の書誌では，書誌項目の定義および用法の理解に文化財についての予備知識が求められる.

さらに，書誌項目に階層がある場合，表形式で階層データを表現するためには，複数の表が必要となる．あるいは，階層化が必要な項目を一つの項目として表現するため，項目あるいは値を非構造化する，つまり単なるテキストとする場合も散見される．その結果，他者が再利用するのが困難な書誌データとなる.

Linked Data は，メタデータのみならず，書誌項目の情報自体を含む構造化データを可能とし，前述の課題を解決することができる．さらに，Linked Data 形式の書誌データを作成しておけば，データベースシステムの維持が困難になったとしても，別のシステムで他者が復旧させやすくなる．また，外部の Linked Data と何らかの関係を定義して元の情報を拡張するなど，データの充実が可能となる.

しかし，書誌を Linked Data 形式に変換するには，書誌の設計図にあたる構造化情報を作成する必要がある．加えて，自然文で記述された部分も

表 1　小城藩日記データベースの記事文に対して設定した固有表現クラス

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

含めた値から，キーワードとなり得る固有表現を抽出し一意の識別子を付与することも必要である.

このような作業は，手作業で行うには膨大であるため，日本語の場合，形態素解析あるいは固有表現抽出支援ツールを用いて可能な限り自動でデータ化を行う．しかしながら，このような抽出ツールの多くは，現代文に対する解析を想定している．そのため，近世の日記目録に記されているが，現代では使用されない候文を対象としたい場合は，別途ユーザ辞書を作成する必要がある.

筆者がこれまで行なって来た固有表現抽出では，表 1 に挙げる Web や書籍などで収集した候文の定型句，人名，地名などを収集し，形態素解析用ユーザ辞書を作成してきた．しかし，今回の対象である小城藩日誌のような文書に含まれる地方特有の固有名詞は，関連書籍でも特定することが困難である.

そこで，本研究では，小城藩日記が作成された小城藩に関する郷土資料を所蔵および調査している小城市立歴史資料館に協力を仰ぎ，固有表現抽出をクラウドソーシングで行うための協力者を募った.

郷土資料の出所に近く，日頃から史料の翻刻や読み解きに深い関心を持つ地元周辺の市民は，土地勘があり，地方特有の氏名に詳しい．また，自分の地元で実際に起こった過去の出来事を知ることにも好奇心が旺盛であると考えられる．そのため，不特定多数から人員を募るよりも，古文書を研究する専門家が選定した翻刻文の意味を理解可能な少数の協力者で，固有表現抽出作業を行なうこととした.

3．クラウドソーシングによる固有表現抽出システム

膨大な作業を細分化し，Web を通じて多くの人々で協働して作業を進めるクラウドソーシングは，人文情報学および図書館情報学の分野にお

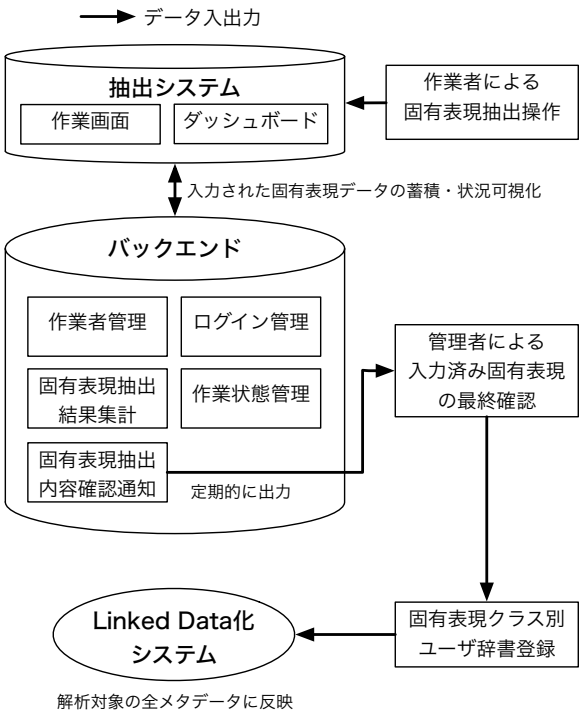

図 2　クラウドソーシングによる固有表現抽出システムの概要および Linked Data 化システムとの関係

いても活用されつつある．自動テキスト化が困難な画像からのテキスト抽出，タグ付けおよび作業自体の見直し，写真中の人物特定，地名抽出，書誌の同定などにおいて，一定の成果を挙げている[8][9][10]．

本論文でのクラウドソーシングでは，固有表現の抽出から分類・登録を容易に行うのみでなく，郷土史料の内容に明るい市民などの協力を得ることができる抽出システムを開発した．さらに，その結果を Linked Data 化システムに固有表現抽出結果を投入することで，書誌中の全メタデータに対しても抽出結果を自動で反映できる．

全体の概要を図 2 に示す．協力者が Web から簡易な操作によって固有表現を抽出し，固有表現クラスに分類・登録する「抽出システム」と，管理者がデータや作業を管理するバックエンドから構成する．

3.1 抽出システムの機能

抽出システムでは，ログイン後，協力者に小城藩日記目録の記事文の一つを作業画面に表示し，固有表現抽出を実行させる．本節では，抽出システムを構成する作業画面およびその進捗状況を表示するダッシュボード画面について説明する．

3.1.1. 作業画面

作業画面は，協力者が翻刻文から固有表現を同定する作業を行う中心となるものである．画面は，

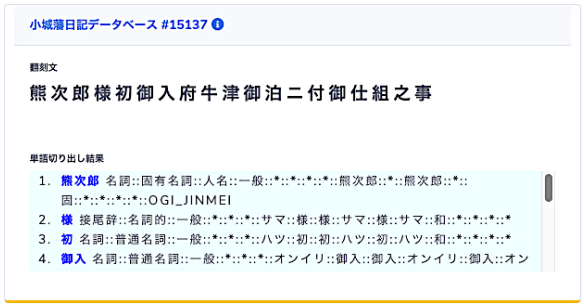

図 3-1　ランダムに選ばれた記事文（翻刻文）および形態素解析（単語切り出し）結果の例

図 3-2　固有表現クラスへの自動ラベリング結果（左）およびその結果の修正機能（右）の例

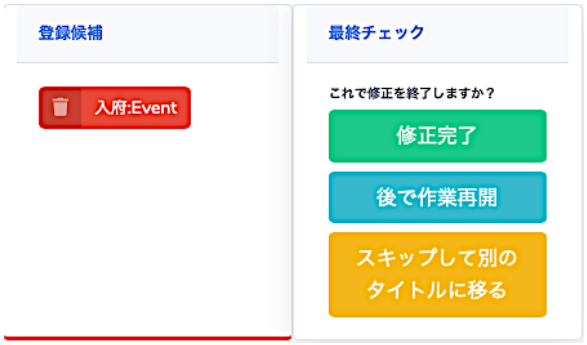

図 3-3　図 3-2 右の固有表現クラス修正機能で新たな登録候補を追加した例

図 3-1，3-2，3-3 および捕捉説明，作業状態，作業履歴を含む 9 個のペイン（枠）で構成されている。

作業者がログインすると，ランダムに選ばれた記事文（翻刻文）およびこの時点で利用できるユーザ辞書を用いた形態素解析（単語切り出し）結果が示される（図 3-1）．また，形態素解析による固有表現抽出の結果が「自動ラベリング結果」として示される（図 3-1 下部及び図 3-2 左）．その際，一つの固有表現に固有表現のクラス名が複

数存在する場合がある. 例えば, 図 3-2 左中の「初」には，「出来事」および「数量，単位」のラベリングが行われている．その場合は，自動ラベリング結果の部分に重複して表示させている．作業者が，この結果を修正する必要があると判断した場合には，「修正」部分に正しい固有表現とそのクラスを入力し，「候補にする」ボタンを押す（図 3-2 右）．その結果は，「登録候補」として図 3-3 左に表示される．更に修正する必要があるものが「自動ラベリング結果」にある場合には，「修正」部分に次の修正内容を入力する．

対象となる記事文に関する修正完了後，「最終チェック」の「修正完了」を押す（図 3-3 右）．作業者は，作業を中断して，後から再開することや，別の記事文へ移動することもできる．

3.1.2. ダッシュボード画面

ダッシュボード画面は，作業者全員が，固有表現抽出作業の進捗を確認する画面である．この画面は，図 4-1，4-2 およびプロジェクトの概要，固有表現集計結果などを含めて 5 個のペインで構成される。図 4-1 の固有表現チェックペインでは，作業者によって確認済みとなった記事文の総数を表示する．図 4-2 の月別進捗状況では，固有表現抽出作業が完了状態の記事文を総計して表示する．

このほかに，作業者に対する共通メッセージと，各作業者が作業を行なった記事文の文字数およびそれに対する謝金の総計と月別合計表示を行うペインがある．

3.1.3. 抽出結果の利用

抽出結果は，固有表現クラスごとにリスト化する「固有表現抽出内容確認通知」スクリプトにより，定期的にファイ四へ出力される（図 2 のバックエンド）．管理者および専門家は，生成された固有表現一覧ファイ四を目視で最終確認する．その後，新規の固有表現を形態素解析用ユーザ辞書に投入する．

固有表現データの投入後，ユーザ辞書の構築および全記事文に対する形態素解析を，図 1 および図 2 に示す Linked Data 化システムによって定期的に自動実行する．

3.1.4. 固有表現クラスへの分類作業

固有表現抽出作業を行う際には，クラスへの仕分けルールの他に，どの固有表現がどのクラスに該当するかを具体的に示した事例集を，作業者が自主的に作成し，共有している．加えて，作業依頼者と作業者間でルールのすり合わせを適宜行なっている．

図 4-1　作業者全員の固有表現抽出作業において，作業完了となった記事文の件数を表示するペインの例

図 4-2　固有表現抽出作業が完了した記事文の総数を月別に集計した結果

図 5　固有表現リストで「肥州」を検索した例. https://winter.ai.is.saga-u.ac.jp/cs/ne-words 参照.

固有表現の抽出とクラス仕分けの結果をLinked Data化システムに投入すると，全記事文の解析結果が更新される．そのため，「固有表現リスト」と呼ぶ，最新の結果を閲覧および検索可能な画面を図5に示すURLに設けた．

固有表現リストで，例えば，地名か人名かクラス判定に迷う固有表現の一例で，「肥州」を検索すると，他の作業者が既に判定済みの語を複数表示する．作業者は，この結果を参考に，クラス判定を行うことができる．

3.2. バックエンドの機能

3.1節および図2で示した抽出システムに必要なデータの蓄積および発行を行い，Linked Data化システムに形態素解析に必要な固有表現に関する情報を出力する，ミドルウェアの役割を果たす機能群を総称してバックエンドと呼ぶ．バックエンドの機能は，大別して作業者管理機能（図2バックエンドの作業者情報管理およびログイン管理）と固有表現集約機能（図2バックエンドの固有表現抽出結果集計，記事文の担当者別作業状態管理および固有表現抽出内容確認通知）に分かれる．

作業者管理機能は，作業者が抽出システムを行うウェブサイトにログインする際に，その状況を時系列のデータとしてデータベースに蓄積し，別途管理者による分析および観察のため可視化する．

固有表現集約機能は，各記事文に対して，作業担当者，日時，作業状態（保留または完了），図3-3で修正した固有表現とそのクラス名の組をデータベースに記録する．さらに，固有表現抽出結果を抽出システム上で作業者にリアルタイムで表示する機能およびクラス別に集約し，形態素解析用辞書データに適用する前段階の確認ファイルを作成する機能を有する．

3.3 現在の状況

抽出システムを利用して固有表現抽出作業を進めるため、江戸時代の業務日誌の候文に詳しく，可能な限り地元の人名や地名に馴染みのある佐賀県小城市の市民を中心に，2019年5月以降，専門知識の提供を行うことを目的に協力者を限定的に募集し，現在8名が参加している．そのうち4名は60代から80代の方であるが，抽出システムの作業画面（図3-1, 3-2, 3-3）での操作は，3.1.1節の通り簡易であるため，事前の簡単な操作作説明だけで，抽出作業を進めている．また，作業進捗をより早く進めるため，希望者5名は有償で，より気軽に作業に参加したい3名は無償で作業を行なっている．

2019年10月の時点で，41719件の翻刻された記事文のうち，約15000件を協力者が修正した．特に，作業者と作業依頼者が顔合わせをし，3.1.2節で述べた研究の概要や細かな固有表現抽出ルールついて具体例を示しながら説明する会を数回行なった7月以降は，作業完了件数が大幅に伸びた（図4-2）．

3.4. クラウドソーシングの品質維持

本抽出システムでは，一人の出した結果がLinked Data化システムによる後の形態素解析結果に大きく影響する．3節に示した，機械的処理の困難な処理を人的資源で行うプロジェクトにおいても，Webで不特定多数に作業を依頼することから，期待する成果が一定期間で得られないという課題がある．Danielらは，クラウドソーシングの品質は，作業協力者の質，作業を行うソフトウェアの質，成果物の質の3つの観点で評価する必要性を示している[11].

抽出システムでの作業協力者は，3.3節に示す通り，候文の文法や用語について十分理解できる程度の経験を有している．作業実態は，アカウント管理ログから時系列可視化ツールを用いて把握することができる（図6）[1]．その結果，1時間

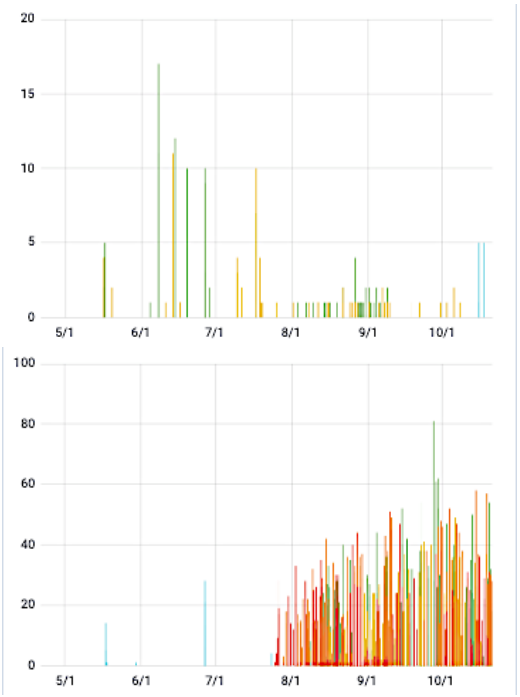

図6　各作業者が1時間あたりに作業完了した記事文の数．（作業者別に色分けしている．）上は無償の作業者，下は有償の作業者．

[1] Grafana, https://grafana.com/grafana/

あたりに 1 件以上の処理がある場合の記事文処理数は, 無償と有償の協力者が, それぞれ 2.5 件, 12.2 件であり, 作業量と速度に明らかな違いが生じた.

ソフトウェアは, ウェブブラウザをマウス操作可能であれば, 参加者全員が問題なく抽出作業ができる程度に簡素にした. 加えて, 説明会の際に, 依頼者や別の参加者同士で一緒に操作を進めることで, 参加者が単独で作業を行う際の不安および疑問点を解消した.

各参加者がどの記事文を担当したか, 図 3-2 の修正作業でどのような固有表現を設定したかをログに蓄積している. 成果物の最終確認は, 作業依頼者で固有表現別データセットを一定期間ごとに目視で確認する. 本抽出システムでは, 各記事文を構成する固有表現が複数のクラスから選択される可能性がある場合, 特定の固有表現がどのクラスを指すのかということまでを正確に特定することは困難である. なぜなら, 固有表現のクラス特定は, 専門家の間で解釈の仕方が分かれる場合があるためである. また, 作業依頼者の決めたクラス判定ルールそのものが, 作業開始当初から普遍的なものではなく, 作業者からの意見でルール修正を行うこともある. そのため, 現時点では, どの固有表現クラスの可能性があるのかを, 図 5 で示す固有表現リストで参照可能とするに留めている. 成果物の質の評価は今後の課題である。

クラウドソーシングによる作業は, 簡易な操作が必要である反面, そのことにより多くの参加者が短期間で興味を失いやすいという性質を持つ [12]. そこで, 抽出した固有表現を, Linked Data を構成するメタデータとし, 小城日記データベース記事文の自動分類や可視化などに利用している[13]. さらに, その結果に対して, 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.

4. まとめ

筆者は, 地域資料の一つである, 小城藩日記のメタデータ化を行う際に, Linked Data 化を行なっている. 本研究では, 手作業では大きな労力を要する Linked Data 化を低コストで行うため, 地域資料に詳しい市民が, 固有表現抽出可能なクラウドソーシングを用いる仕組みを提案した. そのなかで, 適切な人選と作業実態の管理, 簡易なユーザインターフェースを持つ作業画面, 作業依頼者と協力者間での意見交換および協力者のモチベーション維持が作業品質の維持に重要であることを示した. 提案システムは, 資料の Linked Data 化を促進するのみでなく, 市民が自身の文化保護と継承に積極的に参加する方法の一つとなり得る.

謝辞

本研究は, 科研費 19K20630 による助成を受けている.

参考文献

[1] “ジャパンサーチ”. https://jpsearch.go.jp/, (参照 2019-10-17).
[2] “人文学オープンデータ共同利用センター”. http://codh.rois.ac.jp/, (参照 2019-10-17).
[3] “みんなで翻刻”. https://honkoku.org/, (参照 2019-10-17).
[4] “Linked Data: Evolving the Web into a Global Data Space (1st edition)”. Tom Heath and Christian Bizer. Morgan & Claypool, 2011.
[5] “古典籍書誌データ構造に対応した Linked Data への半自動変換”. 吉賀夏子, 只木進一. 情報処理学会論文誌, 59(2), p.257-266, 2018.
[6] “小城藩日記データベース”. https://www.dl.saga-u.ac.jp/ogiNikki/, (参照 2019-10-17).
[7] “小城藩日記データベースの構築”. 吉賀夏子, 只木進一, 伊藤昭弘. 研究報告人文科学とコンピュータ（CH）, 2018-CH-117(3), p.1-7, 2018.
[8] “BY THE PEOPLE”. https://crowd.loc.gov, (参照 2019-10-17).
[9] “Trove Stats for environment: prod”. https://trove.nla.gov.au/system/stats?env=prod#corrections, (参照 2019-10-17).
[10] “Comunidad BNE”. https://comunidad.bne.es/proyectos/categorias/destacados/ (参照 2019-10-17).
[11] “Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions”. Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, Mohammad Allahbakhsh. ACM Computing Surveys, 51(1), p. 1-40, 2018.
[12] “Passerby Crowdsourcing: Workers' Behavior and Data Quality Management”. Eiichi Iwamoto, Masaki Matsubara, Chihiro Ota, Satoshi Nakamura, Tsutomu Terada, Hiroyuki Kitagawa, Atsuyuki Morishima, IMWUT, Vol. 2, p. 169:1-169:20, 2018.
[13] “「今泉惣左衛門」が登場する記事に同時に出現する人名のつながりを可視化した例”. https://www.dl.saga-u.ac.jp/ogiNikki/dashboard2.php?wd=今泉惣左衛門&lm=50&os=0&at=m.date&st=asc&cbe=&sp= (参照 2019-10-17).

