サイエンティフィック・システム研究会 2013 年 10 月 24 日
合同分科会 2013 年度会合 資料

# データ中心時代のメディア～分野を越えた共通認識に向けて

北本 朝展

国立情報学研究所

**[アブストラクト]**

データ中心時代には、データ共有から生まれる共通認識が、社会を動かす基盤になるだろう。データから得られる定量的な現状認識が、人間には避けがたいバイアスを緩和し、よりよい意思決定に結びつくと期待される。しかし最終的な意思決定をするのは人間であり、そこでの情報学の役割は、メディア（媒介）、つまりデータと人間をつなぐ役割にあると考える。そこで、地球環境や防災、文化遺産などの分野で進めてきた研究を「データ中心メディアの創生」というコンセプトでまとめ、その背景にある考え方を紹介したい。

**[キーワード]**

データ中心メディア、地球環境データ、防災データ、文化遺産データ、共通認識

## 1. はじめに

データが根拠となって社会の動きを左右する時代が到来しつつある。しかしそれは必ずしも「ビッグデータ」の時代になることを意味するわけではない。我々はスモールデータ、すなわち量が小さく、種類も少なく、リアルタイム性もそれほど要求されないデータでさえうまく扱えていないが、ビッグかスモールかを問わず、社会の意思決定には良質なデータと良質な解釈が重要な役割を果たすはずだからである。これまでデータに関しては主にアナリティクスの側面が注目を集めてきたが、いくらアナリティクスが進歩してもデータと人間とがきちんと接続されなければ、その結果を社会に活用することはできない。そこで必要となるのがデータと人間をつなぐ「メディア」に関する研究ではないだろうかとの考えに基づき、本発表ではデータをどう見せるか、どう伝えるかといった観点から、発表者がこれまで行ってきた地球環境や防災、文化遺産などのテーマから話題を拾って紹介する。またこのような幅広いテーマを扱うためには、必然的に分野を越えたボーダーレスな研究を進めることにもなる。その過程で発表者が経験してきた問題点などにも触れてみたい。

## 2. エンティティの統合

エンティティとは、実世界の特定のモノを指すものであり、それに固有の ID を付与できれば、これを用いてモノを一意に特定することができる。またエンティティにつけられた名前は固有名（named entity）と呼ばれる。このようなエンティティのレベルでデータを統合することは、分野をまたぐデータ統合において最初のステップとなる処理である。例えば台風に関するあらゆるデータを扱うウェブサイト「デジタル台風」(1)では、気象衛星画像や気象観測データ、マスメディアニュース記事、ソーシャルメディアデータなど大量の異種データを扱っている。これらのデータを統合する際の基本単位となるのは、台風番号という固有の ID であり、それに加えてアメダス観測所や地方自治体、地名など各種のエンティティである。これらを基準としてデータを統合した上で、検索やランキングなどでデータを関連付け、その意味を適切に解釈するための文脈を与えるデータベースを構築するというのが、「デジタル台風」の基本的な設計指針である。

ここで、エンティティがメタデータなどに利用しやすい形で定義されていれば、エンティティを用いたデータ統合には特に問題はない。しかし自然言語テキストなどの形式のデータを扱おうとすると問題が生じる。自然

言語テキストはそのままではマークアップされていないため、まずはどの文字列がエンティティに対応するかを判定せねばならない。こうした処理を特に地名を対象にして行うことを目的としたプロジェクトが「GeoNLP」(2)である。一見すると自然言語テキストから地名を抽出するという処理は簡単に見えるが、実際はそうではない。その根本的な原因はテキストがもともと曖昧であり、複数の解釈が可能であるという点にある。どの部分がエンティティに対応するかを抽出するだけではなく、それがどのエンティティに対応するかを一意に解決するという複雑な処理も必要になってくる。そこで GeoNLP は、このような処理をオープンデータとオープンソースというオープンな環境で実現し、多くの人々がその知見を共有できるようにすることを目指す。またエンティティの統合については GeoLOD という試みも進めている。これは近年注目を集める Linked Data に基づくものであり、Semantic Web 技術を用いてウェブサービスを越えた地名データのリンクを実現する仕組みを提供している。

最後にこうしたアルゴリズムで自動化できないエンティティ統合の例として、遺跡の統合という問題を紹介する。これは我々が進めているデジタルヒューマニティーズ(デジタル人文学)の研究プロジェクトである「ディジタル・シルクロード」(3)で扱う問題である。このプロジェクトで扱っている種々の史料の問題点として、史料によって異なる名前や位置で記載されている遺跡があり、これが原因で昔の史料に出現するのに現在の所在が不明となっている遺跡があった。この問題は史料のテキストを読み込むだけでは解決していなかったが、我々は従来の歴史研究で重視されなかった地図や写真などの非文字史料を適切に読む手段が問題解決に必要であろうと考えた。そこで我々は非文字史料を定量的に扱うことのできる方法論を提案し、遺跡というエンティティをリンクして所在不明遺跡のほとんどを再発見することができた。こうした成功にも関わらず、この方法論は長年にわたって歴史研究者に理解されなかったのが実情である。ところがある時、我々の研究は歴史学の根幹をなす概念である「史料批判」の拡張である、と説明を変えたところ、研究がすんなりと受容されたばかりか、「データ史料批判」は当然存在すべき研究テーマであるとの認識さえ生まれた。この経験は、分野を越えた研究を異分野に正しく伝えるには、その分野の概念を使って伝える必要があることを示していると言える。

## 3. データのストーリー化

このように、分野を越えてデータを共有するためには、データをエンティティのレベルで統合するだけではなく、相手の理解できる形式で伝える努力も必要であろう。これがデータ中心時代のメディアの課題であり、大規模・複雑データを伝える技術についても研究を進める必要があると考えている。その一つの方向性として注目しているのが、データのストーリー化である。例えばジャーナリズムはもともとストーリーを語ることを主要なミッションとしていたため、データからストーリーを語る形式をデータジャーナリズムと呼んで新しい試みを展開している。また最近よく聞かれる用語であるキュレーションも、素材を集めて並べる作業は一種のストーリー作りであるとみなせる。こうした方向の研究として、情報をプッシュ化することで言葉をランダムに遷移させるサイト(311 メモリーズ)(4)、情報をテレビ化することで地域情報を仮想チャンネルのように見せるサイト(311TV)、そして表示メディアを実寸大に拡大することでデータを体感できる展示(伊勢湾台風メモリーズ 2009)などの研究を進めてきた。ただしストーリー化は、行き過ぎると誤ったメッセージを伝える可能性があることにも注意せねばならない。データは一見客観的に見えるため、見る側の批判的精神を失わせる傾向がある。また人間はデータに対してバイアスを伴った判断を下す傾向があるため、データを解釈する文脈を適切に設定してバイアスが小さくなるようにすべきだろう。データを基盤とする社会の実現には、まだまだ様々な工夫が必要である。

## [参考文献]


(1) デジタル台風 http://agora.ex.nii.ac.jp/digital-typhoon/

(2) GeoNLP https://geonlp.ex.nii.ac.jp/

(3) ディジタル・シルクロード http://dsr.nii.ac.jp/

(4) 東日本大震災アーカイブ http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/