

人文学DXを指向する情報基盤の構築

Construction of an Information Infrastructure for the Humanities DX

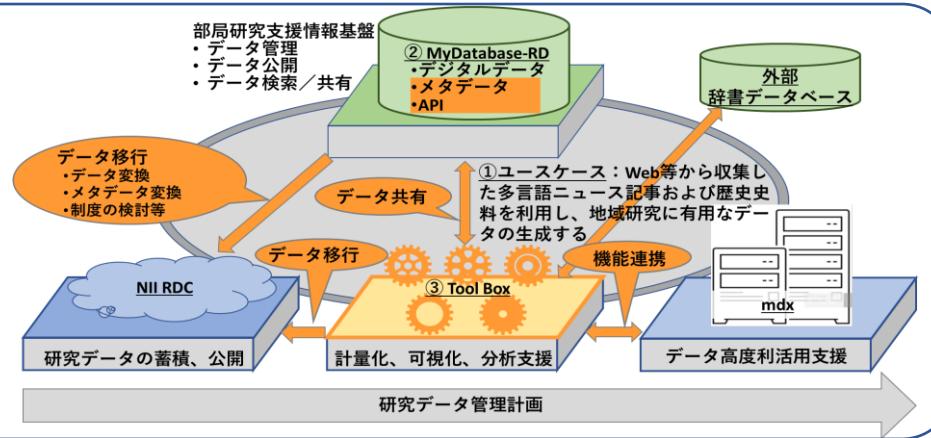
原正一郎^{*1}、馬場弘樹^{*2}、森信介^{*1}、村上勇介^{*1}、関野樹^{*3}、山田太造^{*4}、後藤真^{*5}

Shoichiro Hara^{*1}, Hiroki Baba^{*2}, Shinsuke Mori^{*1}, Yusuke Murakami^{*1}, Tatsuki Sekino^{*3}, Taizo Yamada^{*4}, Makoto Goto^{*5}

1.京都大学(Kyoto University) 2.一橋大学(Hitotsubashi University) 3.国際日本文化研究センター(International Research Center for Japanese Studies)

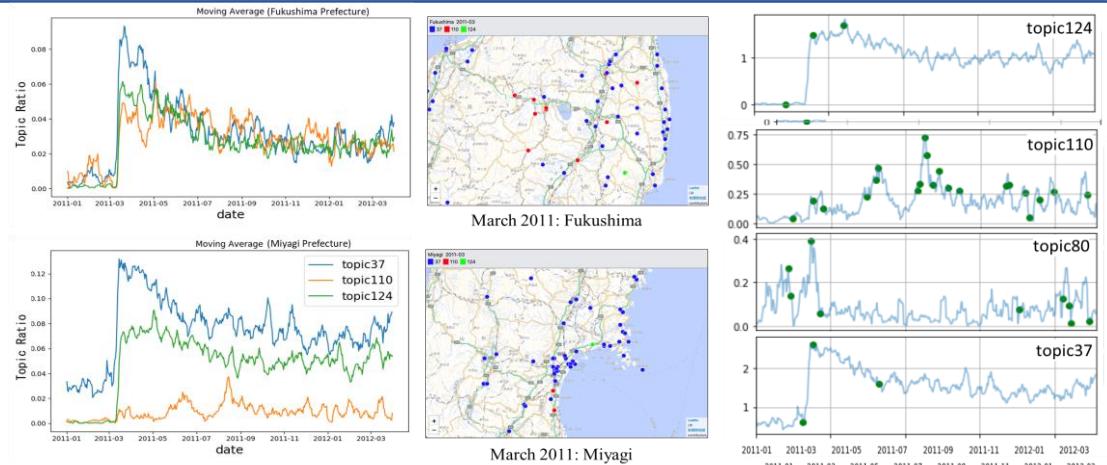
4.東京大学(University of Tokyo) 5.国立歴史民俗博物館(National Museum of Japanese History)

課題の概要：人文学DXの実現には人文学データの計量的処理が不可欠である。そこで本課題では、大量のテキストデータの計量的処理をユースケースとして、①Webニュースデータから地域変動を検出する計量的手法の開発、②データ分析を支援するツールボックスの構築、③関連する大量・多様な人文学研究データの蓄積・共有・利活用を支援するデータベースシステムの構築、および④研究データを適切に管理する手法の確立を試みる。



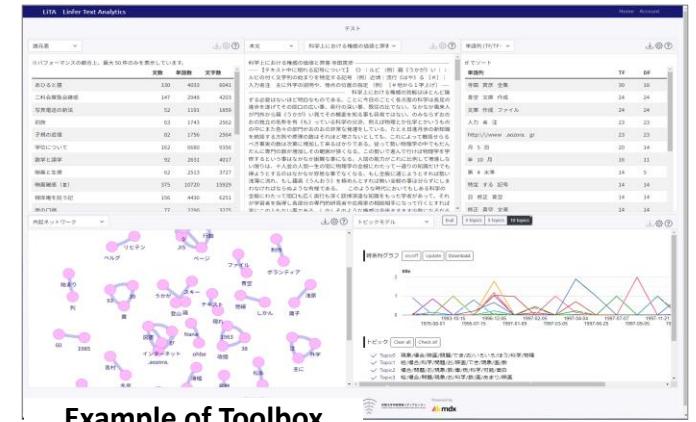
Webニュースデータを利用した地域分析の試み

：ビッグデータを駆使することで、地域社会が変容する瞬間（Tipping Point）を捉えることを目指す。この問いに挑みながら、定性的手法が主流である地域研究において、情報学の計量性を活かした地域研究手法の構築を目的とする。これを実現するために、GakuNin RDMとmdxを利用した以下の課題を実施する。



ツールボックスの開発：テキストデータ分析に必要な機能をツールボックスとして研究者の利用に供する。

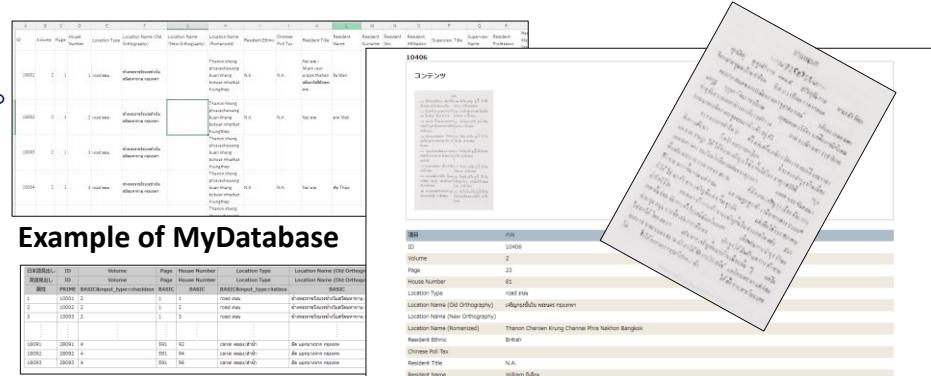
- テキストの語彙分割、時空間語彙認識および計量化（緯度・経度および時間の推定）、主題分析用LDA（Latent Dirichlet Allocation）、基本的な分析結果の可視化に関するツール群を実装した（図）。
- 時系列データ分析や異常点検出等のアルゴリズムの充実を図る。
- 大量テキストデータの学習と予測を実運用に耐えられる時間内で実行するために、mdxあるいは大学スーパーコンピュータの利用を試みる。



データベースの構築：多様な人文学データの保存

と利活用を支援するデータベースを研究者の利用に供する。

- 京都大学東南アジア地域研究研究所が運用しているMyDatabaseを利用する（図）
- 研究データ管理用の簡易的なメタデータを定義した。
- MyDatabaseに研究データ管理機能を追加した。
- MyDatabaseをGakuNin RDM等の外部システムと連携させるAPIの設計する。



研究データ管理手法の確立：よりよい研究

を進める上で必要なデータ管理の標準化を指向する。

- 研究の各段階（計画、データ収集、分析、公開・再利用）においてGakuNin RDMを有効利用する前提となる知識およびスキルの共有・学習・自己評価の仕組みを開発する。
- 京都大学学際融合教育研究推進センターのアカデミックデータ・イノベーションユニットが開発したBasic Rubricとツールの利用を試みる（図）。

