歴史研究におけるデータ利活用事例
—近代日本における府県別平均就学年数の推計—

報告者：山﨑翔平

東京大学大学院経済学研究科
日本学術振興会特別研究員 DC1
yamaSaki.shohei@gmail.com

2018 年 11 月 22 日

概要

概要

自己紹介：山﨑翔平（やまさき　しょうへい）

所属

- 東京大学大学院経済学研究科博士課程
- 日本学術振興会特別研究員 DC1

専攻

- 日本経済史／教育経済学
- そろそろ人文情報学と入れてもよい？

研究

- 近代日本における人的資本形成
  - 学校制度の導入は日本の経済発展にどのような役割を果たしたのか
  - 歴史統計データを使用した計量分析による実証（数量経済史）

概要

大学院生を中心とした
デジタル・ヒストリーの取り組み

Tokyo Digital History

Tokyo Digital History

- 史学系院生を中心に、アーキビスト・エンジニアと協働してDigital Historyを実践し、
  Digital Historyの方法論（研究・教育）を模索するコミュニティ (2017.9～)
  （※日本史2人・東洋史1人・西洋史6人；アーキビスト2人；エンジニア2人；教員1人）
- オンライン史資料を活用する際の留意点を歴史家の間で議論し、共有する
- メンバーひとりひとりが自分のプロジェクトを進め、多分野的な活動を展開
- Python、TEIを基礎スキルとしながら、個々の能力を掛け合わせ、グループで成果を生み出す

→活用可能な技術の範囲を広げ、
テンポ良く、着実に成果を生み出す（※業績一覧も参照）

活動成果と関連動向

学生奨励賞を受賞
第117回
人文科学とコンピュータ研究会

開催報告@東大機関リポジトリ

4月にシンポジウム開催

シンポジウムの発表を発展させて
国際学会でのパネル報告

5月19日の日経新聞に紹介記事掲載

ウェブマガジン
『人文情報学月報』での
連載開始
（2018年5月号～）

関西デジタル・ヒストリー
研究会も菊池信彦氏が企画中

渋沢栄一記念財団デジタル・キュレーター

金甫榮 | アーカイブズ学

デジタル・アーカイブの多義性 / アーカイブズ理論

デジタル時代に史料とどう向き合うか

東大日本史D3

福田真人 | 近代日本貨幣史

公文録 / 史料群の階層 / Webスクレイピング

巨大な史料群のデータを一括入手する

東大西洋史D3

纓田宗紀 | 中世教皇史

Regesta Imperii / 年表・地図

データの活用から公開までを展望する

東大経済史D3

山崎翔平 | 近代日本経済史

府県パネルデータ / Python / Stata

データ加工の再現性を担保する

東大西洋史D3・歴博研究協力者

小風尚樹 | 近代イギリス外交史

延喜式 / データベース構築 / TEI

デジタル技術で分野を越境する

東大西洋史M2

小川潤 | 古代ローマ属州史

Perseus Digital Library / 古典語テキスト解析

テキスト群から語の使用傾向を分析する

お茶大西洋史D1

山王綾乃 | 近世フランスアカデミー史

会員名簿 / データ可視化 / Tableau

統計データの表現方法を探索する

東大西洋史M2

小林拓実 | 近代フランス移民史

Indicateur Marseillais / GIS / CCライセンス

歴史地図にデータを可視化する

提携組織・プロジェクト

国立歴史民俗博物館
総合資料学の創成
INTEGRATED STUDIES OF CULTURAL AND RESEARCH RESOURCES

公益財団法人
渋沢栄一記念財団
Shibusawa Eiichi Memorial Foundation

『百科全書』
啓蒙研究会

石見銀山
世界遺産センター

後援

人文情報学拠点
Digital Humanities Initiative
東京大学大学院人文社会系研究科　次世代人文学開発センター　創成部門
Center for Evolving Humanities, Graduate School of Humanities and Sociology, The University of Tokyo

国立歴史民俗博物館
総合資料学の創成
INTEGRATED STUDIES OF CULTURAL AND RESEARCH RESOURCES

図書出版
文学通信
Bungaku-Report.com

概要

平均就学年数（Average Years of Schooling）

教育の普及度を測る指数

- ある社会における人々が平均して何年の学校教育を受けているか
- 各年・各コホートにおける就学者数を積み重ねることで推計

加工可能性に富む

- 男女別に
- 教育段階別に（初等、中等、高等教育）
- 教育種別に（普通教育、実業教育）

Barro and Lee (2013) のデータセットが有名

- 各国について第二次大戦後から 5 年おきに推計
- `http://www.barrolee.com/`

近代日本における府県別平均就学年数の推計

史料：文部省年報（第 1 巻〜第 63 巻）

- 複線型と呼ばれる複雑な教育体系を整理
- 各年・各学校の府県別に就学者数をひたすら入力
- 県境が現在と異なる時期については郡の人口情報で補正

府県別人口情報（別に推計）を使用して以下の式により推計

$$AYS_{i,t} = \frac{\sum_{s=1873}^{t-5} E_{i,s}}{L_{i,t}}$$

- ここで、$AYS_{i,t}$ は t 年における i 県の平均就学年数、$E_{i,s}$ は s 年における i 県の就学者数、$L_{i,t}$ は t 年における i 県の人口

今のところ初等普通教育（尋常小学校・高等小学校）について推計

府県別平均就学年数

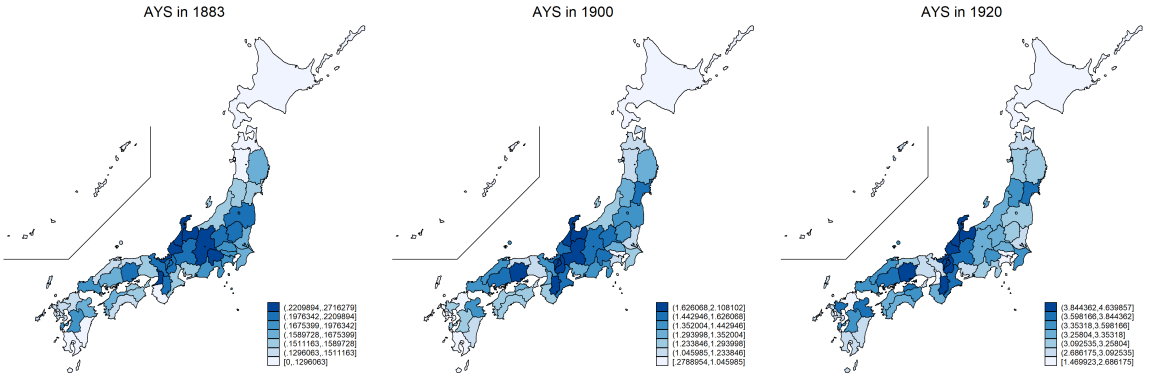
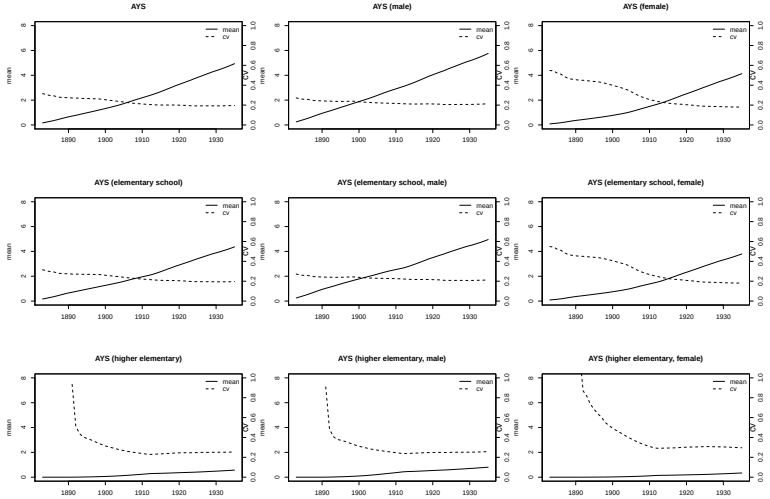


Figure: Painted maps of average years of schooling by prefecture

使用例：平均就学年数 (府県平均と変動係数の推移)



概要

歴史研究と OCR（文字認識）技術

歴史研究

- 紙ベースの史料を文字データとして多かれ少なかれ電子化することが不可欠
- 特に統計学を用いた計量分析には大量のデータが必要

現状：膨大な手間と人件費をかけた手入力

一方で、歴史史料の画像データの公開が進んできた

既存の OCR ソフト

- 文章の認識精度は高いが、統計表は難しい
- OCR の結果から必要な箇所を取り出して、整形するのが手間

ないなら自分で作ってしまおう

画像の必要な箇所を選択（Crop）して、それを OCR にかける

- リソース（通信量、処理能力）の節約
- 認識精度の向上
- 整形の手間が少なくなる

大量のデータを入力することに最適化された UI（操作方法）

- 競合アプリ 1：スマホカメラで撮影
- 競合アプリ 2：スクリーンショットを取得してから選択

誰にも真似できないデータの量と質によって、既存の研究の限界を乗り越えることができる（かもしれない）

Ocrop 歴史研究のための OCR 補助アプリ

Ocrop（オクロップ）

- OCR + Crop

Google Cloud Vision API を使用

- Crop した画像を API に渡す

選択、整形、貼付けの過程をシームレスに統合した UI

- 実際に入力作業をしてきた開発者のこだわり

反響

研究室の同僚による試用の感想

- 縦書き・旧字体も認識できてすごい！
- 固有名詞（人名・地名・会社名）の入力は捗りそう！
- RA の時給が 1000 円だとして、月額 500 〜 1000 円なら払う

『東京大学 150 年史』編纂準備作業に採用

- 卒業生名簿の電子化に使用

展望

IIIF ビュワーへの組み込み

- IIIF（トリプルアイエフ）：画像資料公開のための国際規格
- 主要な IIIF ビュワーではプラグインをサポート

OCR 時の前処理のノウハウの蓄積

- どのように切り出すと認識精度が高いか
- ノイズ除去や鮮明化処理

機械学習による切り出し作業の自動化

- 表の罫線や余白を認識させる

より大量のデータをより短時間で電子化する

まとめ

研究者は喉から手が出るほどデータが欲しい

- データの質は研究の質に直結する
- 入手・加工の苦労は厭わない

どんな形の公開でも研究者は使用する

- 検索できるだけでもありがたい
- 画像ならなおさら
  - Ocrop の試用者や共同開発者を探しています
- 全文検索できれば最高
- 欲を言えば TEI や IIIF への対応を

謝辞

本研究は JSPS 科研費 16J06613 の助成を受けたものである。

Ocrop の開発に際し東京大学 Summer Founders Program 2018 の支援を受けた。

データ入力作業に尽力して頂いた RA 諸氏に感謝したい。