歴史研究におけるデータ利活用事例
ー公文録のWebスクレイピングー

東京大学大学院　日本史学研究室
学術振興会特別研究員（ＤＣ1）
福田真人

主な内容

- 史料とデータベースの紹介
- Webスクレイピングによる情報の入手
- 入手したメタデータの分析

Webスクレイピング

日本研究では相対的に不足しているものの、
オンライン上の史料情報が増加してきている
　→こうしたデータの入手・活用が課題

Webスクレイピング
　ウェブサイトからデータを機械的に集める技術
　人間では不可能な大量の情報の入手が可能
⇒これらのデータを分析する

史料の紹介

・明治太政官期（明治元・1868～明治18・1885）

太政官：当時の最高行政機関

諸官庁などの上申を決裁

・公文録　明治前期の太政官の公文書類

原文書に近い貴重な史料（1873 皇城炎上）

「政府記録ノ基礎」とされる

史料群レベル（公文録）――簿冊レベル（約4000冊）

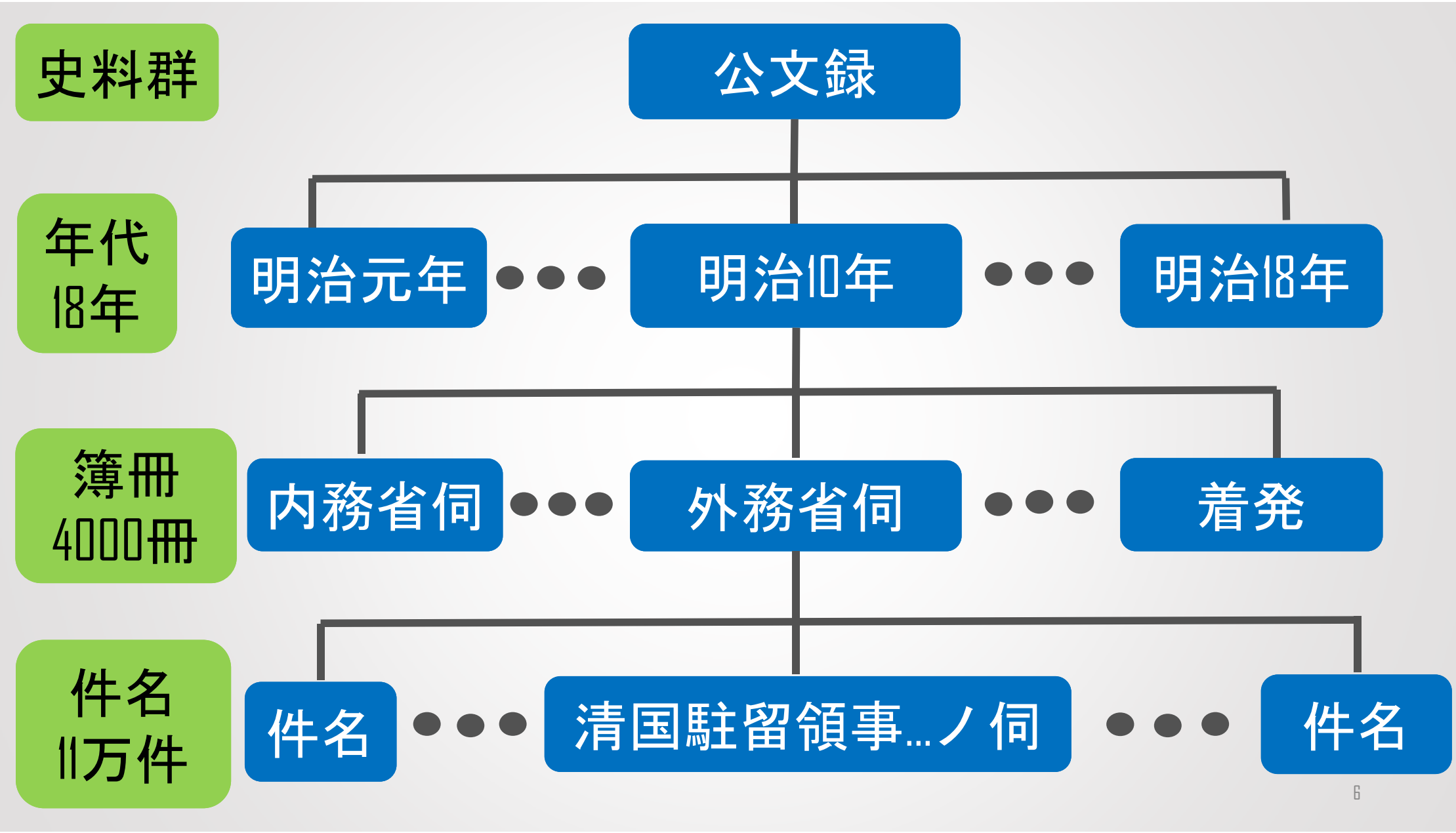
――件名レベル（約11万件）

データベースの紹介

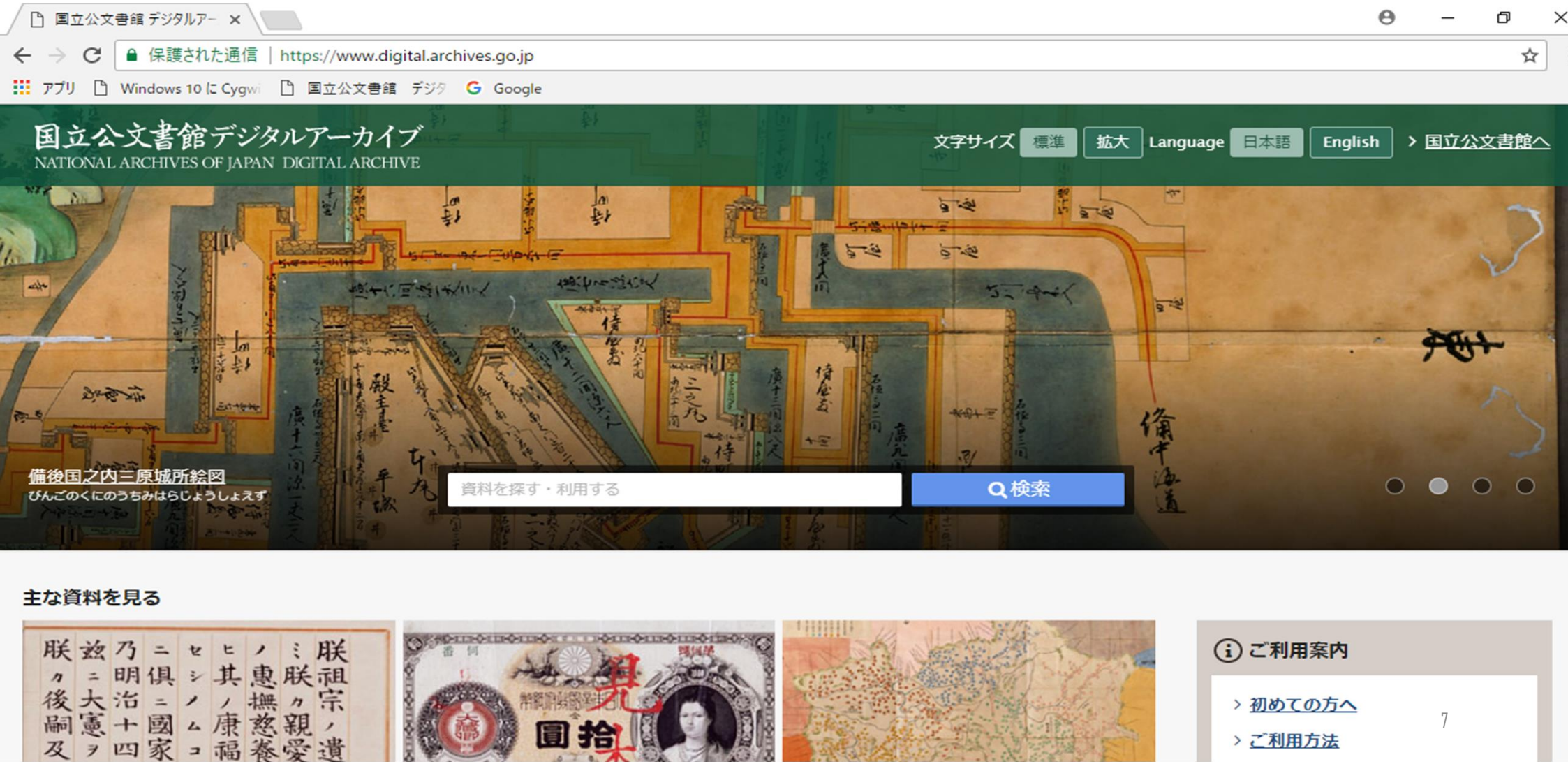
- 国立公文書館　デジタルアーカイブ

  日本を代表するアーカイブスの一つ

  階層的構造


史料群
公文録
年代
18年
明治元年
明治10年
明治18年
簿冊
4000冊
内務省伺
外務省伺
着発
件名
11万件
件名
清国駐留領事…ノ伺
件名

国立公文書館デジタルアーカイブ トップページ



公文録　トップページ


国立公文書館 デジタルア
保護された通信 | https://www.digital.archives.go.jp/DAS/meta/Fonds_F2005032421074303276
アプリ
Windows 10 に Cygwi
国立公文書館 デジタ
Google
内閣人事局文
第一類　雑種公文
第一類　公文録
公文録・明治元年
公文録・明治２年
公文録・明治３年
公文録・明治４年
公文録・明治５年
公文録・明治６年
公文録・明治７年
公文録・明治８年
公文録・明治９年
公文録・明治１０年
公文録・明治１１年
公文録・明治１２年
公文録・明治１３年
資料群階層を閉じる
資料群情報
前の画面へ戻る
検索条件を表示
資料群詳細を表示
タイトル
第一類　公文録
階層
行政文書 > ＊内閣・総理府 > 太政官・内閣関係
年月日
１８６８（明治元）－１８８５（明治１８）
作成部局
〔内閣記録保存部局〕
内容
太政官において授受した公文書を年別・各省庁別に編集したもの。「公文附属の図」，「…
資料群一覧
該当件数: 18件（１－18件）
表示順
昇順
指定無し
一覧の表示内容
標準
切替
No
概要情報
1
公文録・明治元年
2
公文録・明治２年
3
公文録・明治３年

公文録　年代別ページ


国立公文書館 デジタルア
保護された通信 | https://www.digital.archives.go.jp/DAS/meta/MetSearch.cgi?DEF_XSL=default&IS_KIND=summary_normal&IS_SCH=META&IS_STYLE=default&IS_TYPE=meta...
アプリ
Windows 10 に Cygwi
国立公文書館 デジタ
Google
公文録・明治８年
公文録・明治９年
公文録・明治１０年
公文録・明治１１年
公文録・明治１２年
公文録・明治１３年
公文録・明治１４年
公文録・明治１５年
公文録・明治１６年
公文録・明治１７年
公文録・明治１８年
第一類　公文録（副本）
第一類　公文附属の図
第一類　公文附属の表
第一類　公文別録
第一類　公文雑纂
資料群階層を閉じる
資料群情報
前の画面へ戻る
検索条件を表示
資料群詳細を表示
タイトル
公文録・明治１０年
階層
行政文書 > ＊内閣・総理府 > 太政官・内閣関係 > 第一類　公文録
簿冊一覧　該当件数: 230件（１ – 100件）
1 2 3 次 > 最後
画像閲覧可能なデータ
利用請求書印刷
表示順 昇順 指定無し 一覧の表示内容 標準 切替
利用請求
No
概要情報
利用制限区分等
画像等
公文録・明治十年・第一巻・明治十年一月・寮局伺（本局～大使事務掛）
[請求番号] 公02008100 [保存場所] 本館 [作成部局] 太政官 [年月日] 明治10年01月-明治10年[マイクロフィルム]
<件名一覧があります>
1
公開
原本閲覧（否）
閲覧
公文録・明治十年・第二巻・明治十年二月・寮局伺（法制・調査・修史館・式部寮）

公文録　簿冊別ページ


国立公文書館 デジタルア
保護された通信 | https://www.digital.archives.go.jp/DAS/meta/MetSearch.cgi?DEF_XSL=default&IS_KIND=summary_normal&IS_SCH=META&IS_STYLE=default&IS_TYPE=meta...
アプリ
Windows 10 に Cygwi
国立公文書館 デジタ
Google
公文録・明治８年
公文録・明治９年
公文録・明治１０年
公文録・明治１１年
公文録・明治１２年
公文録・明治１３年
公文録・明治１４年
公文録・明治１５年
公文録・明治１６年
公文録・明治１７年
公文録・明治１８年
第一類　公文録（副本）
第一類　公文附属の図
第一類　公文附属の表
第一類　公文別録
第一類　公文雑纂
資料群階層を閉じる
簿冊標題
公文録・明治十年・第十五巻・明治十年八月～九月・外務省伺（八月・九月）
階層
行政文書 > ＊内閣・総理府 > 太政官・内閣関係 > 第一類　公文録 > 公文録・明治１０年
請求番号
公02022100
保存場所
本館-2A-010-00
作成部局
太政官
年月日
明治10年08月 - 明治10年09月
利用制限区分
公開
原本の閲覧
否
マイクロフィルム
リール番号：025500、開始コマ：0370
画像データ
閲覧
画像一括ダウンロード
件名・細目一覧　該当件数: 28件（1 - 28件）
画像閲覧可能なデータ

公文録　件名別ページ


国立公文書館　デジタルア
https://www.digital.archives.go.jp/DAS/meta/MetSearch.cgi?DEF_XSL=default&IS_KIND=detail&IS_SCH=META&IS_STYLE=default&IS_TYPE=meta&DB_ID=G9100001EXTERNAL...
アプリ
Windows 10 に Cygwi
国立公文書館　デジタ
Google
公文録・明治８年
公文録・明治９年
公文録・明治１０年
公文録・明治１１年
公文録・明治１２年
公文録・明治１３年
公文録・明治１４年
公文録・明治１５年
公文録・明治１６年
公文録・明治１７年
公文録・明治１８年
第一類　公文録（副本）
第一類　公文附属の図
第一類　公文附属の表
第一類　公文別録
第一類　公文雑纂
資料群階層を閉じる
件名・細目詳細
前の画面へ戻る
件名/細目一覧へ
件名
清国駐留領事井朝鮮国駐留管理官へ裁判権限委任ノ儀伺
階層
行政文書 > ＊内閣・総理府 > 太政官・内閣関係 > 第一類　公文録
> 公文録・明治１０年
> 公文録・明治十年・第十五巻・明治十年八月～九月・外務省伺（八月・九月）
請求番号
公02022100
件名番号
010
保存場所
本館-2A-010-00
作成部局
太政官
年月日
明治10年08月
受入方法
移管
媒体の種別
紙
利用制限区分
公開
原本の閲覧
否

公文録　簿冊別画像

公文録・明治十年・第十五巻・明治十年八月～九月・外務省伺（八月・九月） - Google Chrome

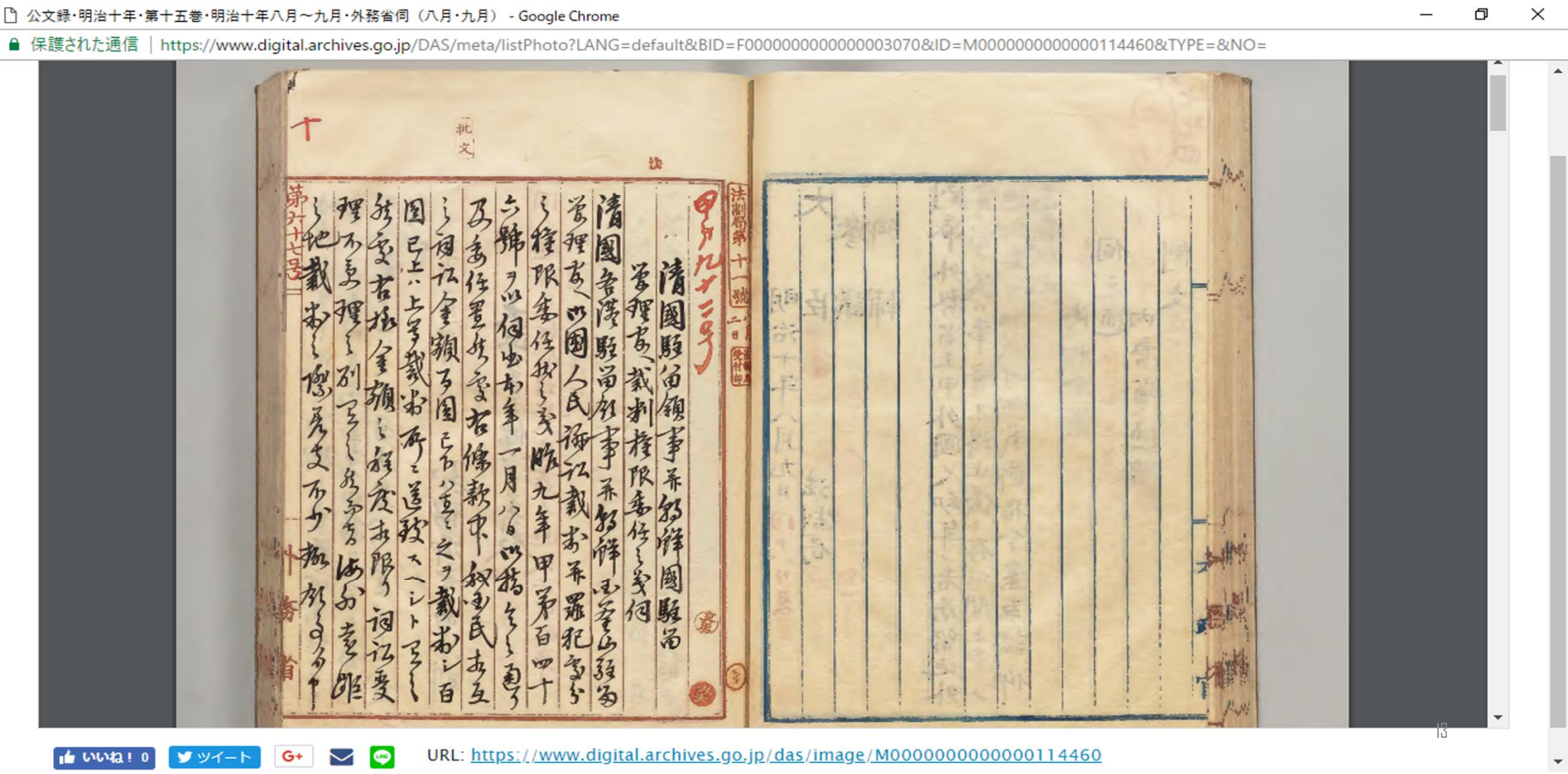
保護された通信 | https://www.digital.archives.go.jp/DAS/meta/listPhoto?LANG=default&BID=F0000000000000003070&ID=&TYPE=&NO=

簿冊標題：公文録・明治十年・第十五巻・明治十年八月～九月・外務省伺（...

PDF　JPEG　画像一括ダウンロード

次の件名 ›　末尾 »　簿冊表紙・目次等

公文録　件名別画像


公文録・明治十年・第十五巻・明治十年八月～九月・外務省伺（八月・九月） - Google Chrome
保護された通信
https://www.digital.archives.go.jp/DAS/meta/listPhoto?LANG=default&BID=F0000000000000003070&ID=M0000000000000114460&TYPE=&NO=
URL: https://www.digital.archives.go.jp/das/image/M0000000000000114460

先行研究と課題

通常の研究はテーマ・省庁別に収斂

全体を論じる近代史料学の研究は不足

先行研究は事例分析が中心

全体分析は冊数のみ（＝人力での限界）

件名ベースでの悉皆調査を実施する必要

→記述統計的あるいは計量的分析

省庁横断的分析・全体における位置づけ

本メタデータの分析意義

- 「メタデータ」であるという限界
- 日本近代史研究において
  重要かつ巨大な史料群
- メタデータは全体にわたって
  史料群内部の詳細な目録（件名）がある
- 目録は史料原本の目次に基づいており、
  後年の作為が少ない

成果

・件名・年代・簿冊名などのメタデータを収集

・10万8840件のメタデータを入手

（データクリーニングを経て）

・史料階層の可視化

・件名の計量文献学的分析などが可能に

階層的バブルチャート

＊小風尚樹氏と共同で作成した

https://bl.ocks.org/naoki-kokaze

Cf) 内務省の件数

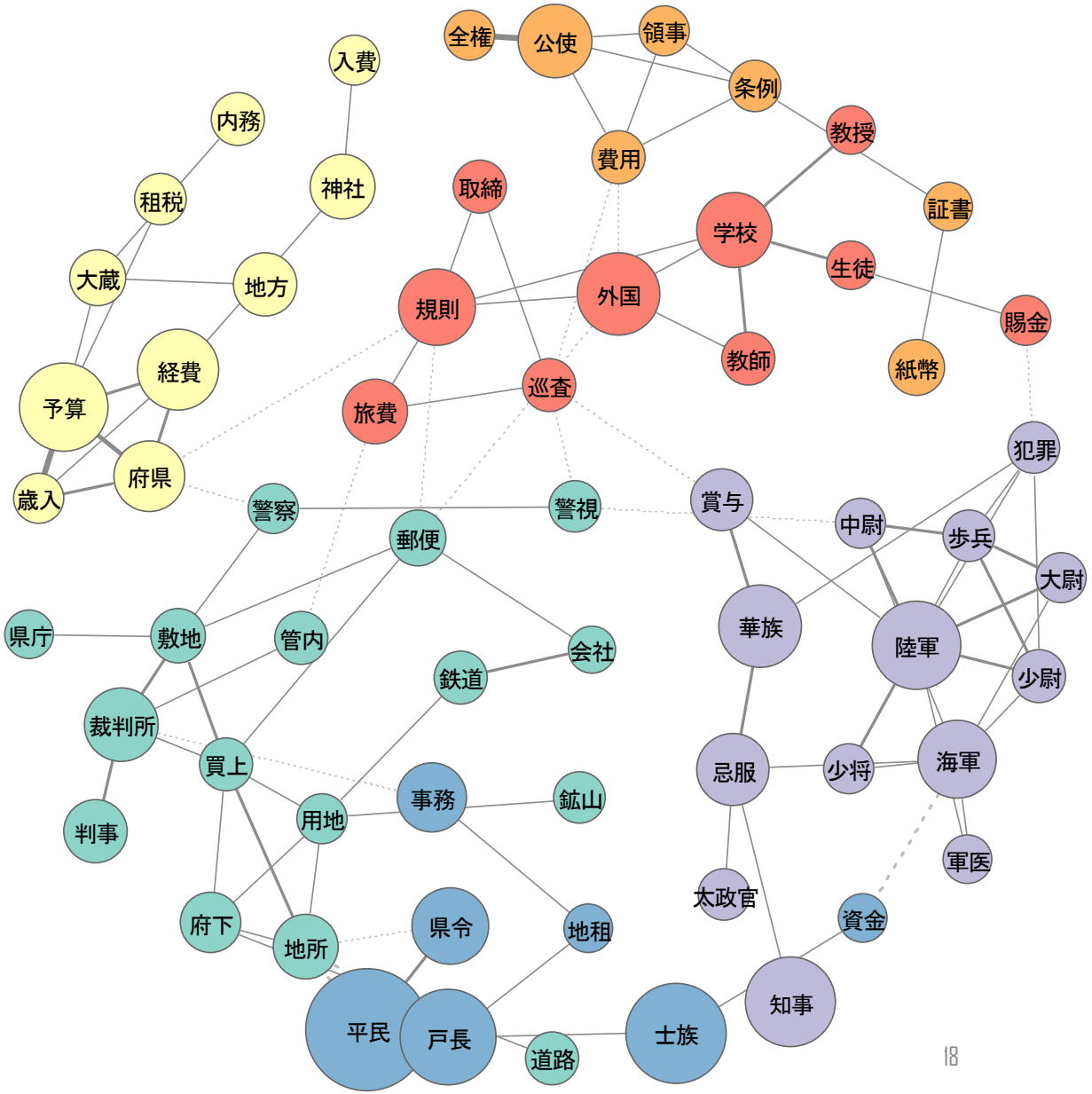
大久保利通暗殺（明治11年）前の時期が最大

内務省時代（～明治14年）に

確かに内務省の件数が多い

⇒通説的イメージを支持しうる定量的データ

公文録件名共起ネットワーク



経済財政策史における明治14年

明治14年の政変（8.31大隈追放決定　10.12免官）

大隈財政⇒松方財政（10.21大蔵卿就任）

積極政策から緊縮政策へ

研究史上の論争

①14年の政策担当者変化を重視

②末期大隈財政（13年）による緊縮を画期とみる

計量文献学に見る経済政策の転調

単語の出現数はその政策の課題規模と類似と仮定

⇒実際には問題があり、今後の課題

明治14年以前を大隈財政期、

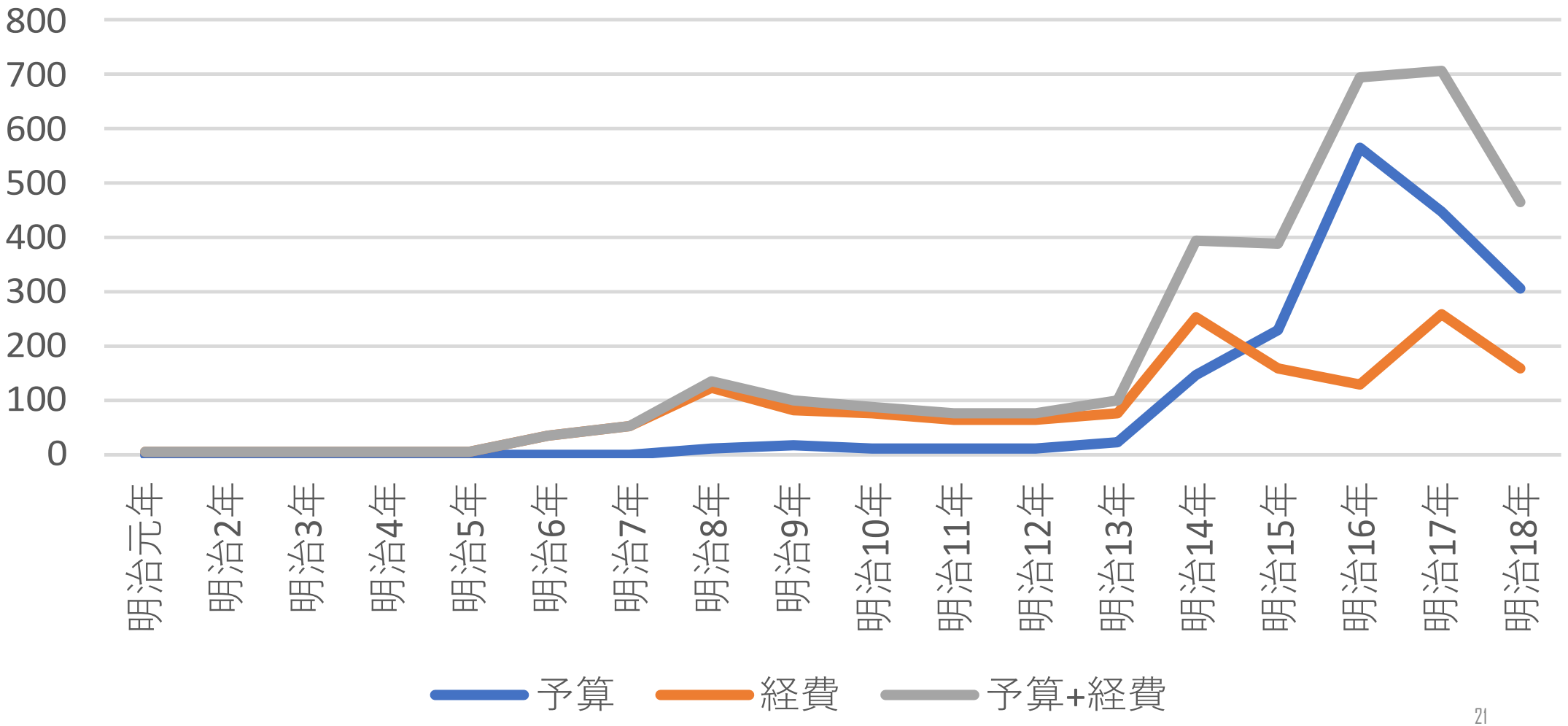
明治15年以降を松方財政期とする

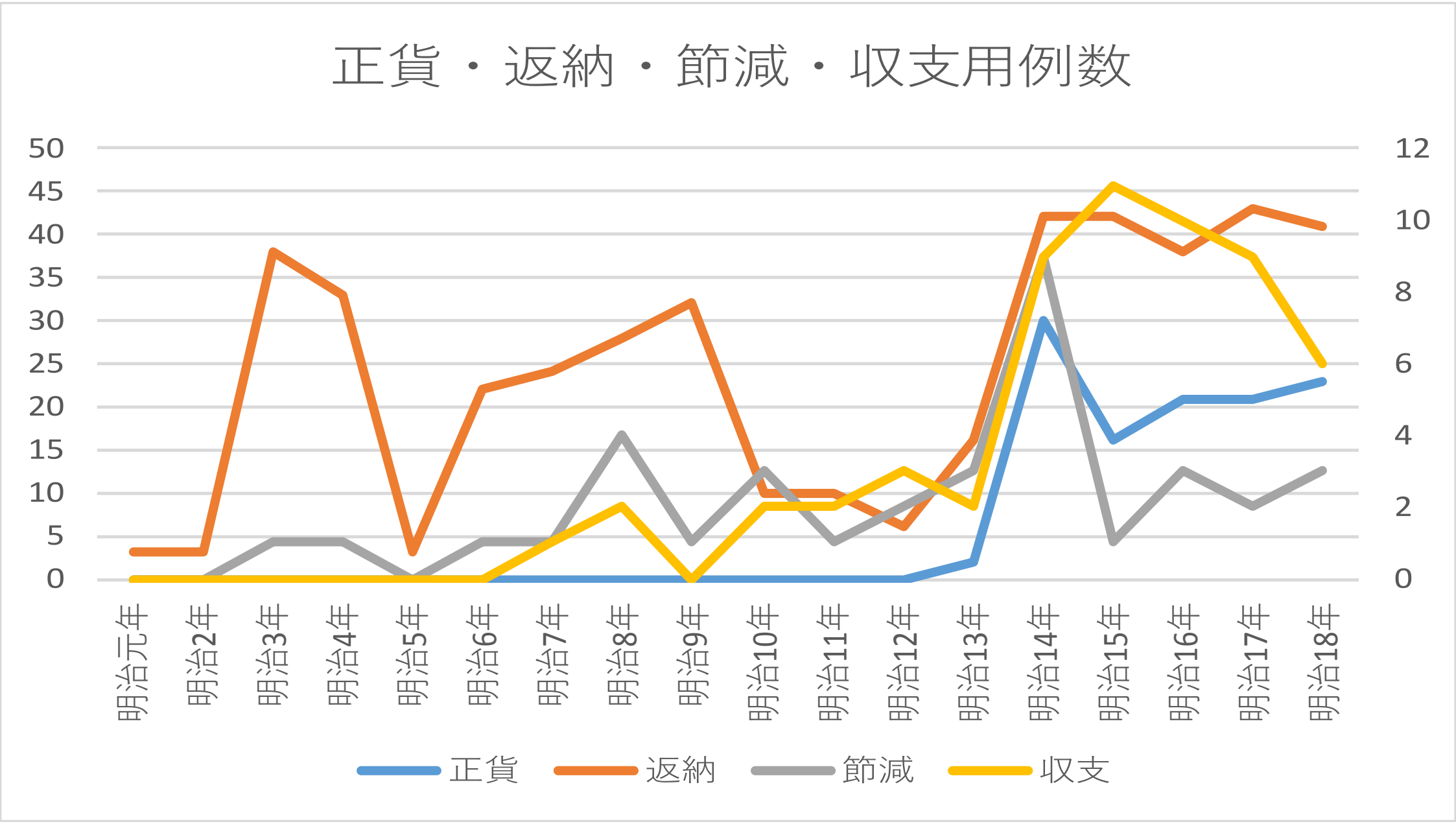
（厳密に月別にしても同傾向　但し決裁日基準）

⇒財政担当者の変化前にキーワードの傾向変化

さらなる文献史学的検討が必要なことは無論

予算・経費用例数


800
700
600
500
400
300
200
100
0
明治元年
明治2年
明治3年
明治4年
明治5年
明治6年
明治7年
明治8年
明治9年
明治10年
明治11年
明治12年
明治13年
明治14年
明治15年
明治16年
明治17年
明治18年
予算
経費
予算+経費


正貨・返納・節減・収支用例数
50
45
40
35
30
25
20
15
10
5
0
12
10
8
6
4
2
0
明治元年
明治2年
明治3年
明治4年
明治5年
明治6年
明治7年
明治8年
明治9年
明治10年
明治11年
明治12年
明治13年
明治14年
明治15年
明治16年
明治17年
明治18年
正貨
返納
節減
収支

件名メタデータの正確性①

・①「公文録」の目次自体が正確か

⇒明らかな誤りは管見の限り僅少

解釈や表現の問題は常に残存する

・②本メタデータは目次を正確に反映しているか

⇒管見の限りミスは少なくない

無作為抽出した100件の件名について

史料群の目次とメタデータの対比を行った

表記ゆれを除くと誤っていた事例は7例

件名メタデータの正確性②

「佐伯郡草津村沖外一ヶ所開墾并見合届」の「届」→「留」

「華族奥平昌邁忌服届遠慮届」の「忌服届」不要

「参事院議官西園寺公望病気ニ付西京へ換地療養願並出発ノ件其三」の「其」→「共」

「東京開成学校教師仏人レビシユ増給雇継并同所へ仏語学教師一名雇入ノ儀伺」の「ユ」→「エ」

「戸籍表差出ニ付申立」の「立」→「出」

「各国人へ対シ凶状無之様御布告藩邸内外へ告諭届」の「凶」→「乏」

「太政大臣三条実美帰京ノ件其二」の「其」→「共」

本メタデータの正確性③

分析対象たりうると現時点では評価しているが、
単純に無批判に採用できるとは言えない水準
単純な誤植の類が中心であることに鑑みれば、
文字率ベースでみる限りは、
統計的にはほぼ無視しうるか？

まとめ①

・デジタルヒストリーの時代へ

デジタル時代特有の情報入手法によって

既存の研究が行えなかった

or発想がなかった研究が可能に

・デジタルヒストリーが発達している欧米に比べると

日本研究においては具体的研究自体僅少

⇔逆説的に研究のフロンティア

まとめ②

・文献史学などの研究との併用が望ましいことは自明

必ずしも既存の研究と対立するものではない

・デジタル特有の気を付けなければならない論点

それを具体化したうえで極力解決することが課題

・デジタル時代の共同研究の重要性