

タイトル

NDLデジタルライブラリーカフェ「2021年の人文学」(20210115)

TEIを通じた粒度の細かいテキスト検索

講師

千葉大学 人文社会科学系教育研究機構 助教

小風尚樹

1. 自己紹介

- 2016年4月～
 - 東京大学 西洋史学専門分野 博士課程
 - 20世紀転換期イギリス海事史
 - 企業の財務・会計史料の数量的分析に、今回紹介するデジタル技術を活用
- 2020年4月～
 - 千葉大学 人文社会科学系教育研究機構 助教
 - デジタル・ヒューマニティーズ(DH)の大学院教育
- 2016年9月～
 - 国立歴史民俗博物館(歴博) 共同研究員
 - とくに『延喜式』のデジタル校訂テキストの作成
 - cf. 小風尚樹・後藤真「『延喜式』へのTEI適用と日本史資料のテキストデータ共有・流通」『国立歴史民俗博物館研究報告』第218巻、315-327頁、2019年12月

2. 『延喜式』とは

- 奈良～平安時代にかけての律令体制下の施行細則(行政マニュアル)
 - 儀礼や祭式
 - 律令政府への租税の詳細
 - 各国ごとの財源の詳細 などなど
- 『延喜式』を通じて古代日本の諸相が窺える
- 右画像出典: 国立国会図書館デジタルコレクション「延喜式. 第1」
<https://dl.ndl.go.jp/info:ndljp/pid/1273518>

祭日	祭記	大 中 小
四年祭上。	四時祭上。	
二月祭。	凡踐 祚大嘗祭爲大祀。新年。月次。神嘗。新嘗。賀茂等祭爲中祀。大忌。風神。鎮花。三枝。相嘗。鎮魂。鎮火。道饗。園韓神。松尾。平野。春日。大原野等祭爲小祀。風神祭已上並諸司齋之。鎮花祭已下祭官齋之。俱小祀祭官齋者内裏不齋其道勅使之祭者齋之。	
新年祭神。三千一百三十二座。	凡新嘗祭二月四日。大忌風神祭並四月七月十四日。月次祭六月。十二月十一日。神嘗祭九月十一日。其子午卯酉等日祭各載本條。自餘祭不定日者。臨時擇日祭之。	
大四百九十二座。三百四座。案上官幣。		
小二千六百四十座。四百卅三座。案上官幣。		
二千二百七座。國司所祭。		

延喜式 卷一 神祇一 四時祭上

一九

3. 延喜式プロジェクトの紹介

- 歴博主導「古代の百科全書『延喜式』の多分野協働研究」
 - 国文学研究資料館「異分野融合による「総合書物学」の構築」の一部
- 1. 古代の知識と技術の現代的活用に向けた分野横断型研究
 - 薬学・食品学・考古学などとの連携
 - 古代日本における貢納品の再現など(例:鮫鮓)
- 2. 海外など幅広い研究者や一般市民に向けた情報発信
 - 英訳・現代語訳のデータベース
 - 写本画像と校訂テキストを比較対照できるデータベースの作成

4. 研究の目的

- これまでの延喜式研究の問題点として指摘されてきたこと
 - 『延喜式』からは、非常に多種多様な情報を得ることができるゆえに、研究対象があまりに細分化されてしまう点
 - 『延喜式』のテキストに基づいて立証された説(とくに租税や貢納品の額・数量)の妥当性を検証することが困難であること
- 古代日本史研究者と情報学およびDH研究者の協働により細やかな検索が可能なテキストデータベースを作成し、様々な先行研究の成果を検証できるようにすること

5. 対策 | テキストのマークアップ

5.1. テキストは情報の宝庫

書面に現れやすいもの

- レイアウト: 見出し・書式・寸法...
- 固有表現: 人名・地名・イベント名...
- 文字装飾: 下線・斜字体・取消線...
- 参照情報: 引用文・出典の記載...
- 文章: 本文(散文・韻文・数値)・註釈...

書面に現れにくいもの

- 書誌情報: 著者名・編者名...
- 来歴情報: 作成・伝承・所蔵...
- 利用条件: ライセンス規定・制限事項...
- 校勘情報: 版ごとの異同

新年	祭日	見出し
祈年祭神。三千一百三十二座。 大四百九十二座。三百四座。案上官幣。 小二千六百四十座。四百卅三座。案上官幣。	二月祭。 日。神嘗祭九月十一日。其子午卯酉等日祭各載本條。自餘祭不定日者。臨時擇日祭之。	四時祭上。 凡踐 昨大嘗祭爲大祀。新年。月次。神嘗。新嘗。賀茂等祭爲中祀。大忌。風神。鎮花。三枝。相嘗。鎮魂。鎮火。道饗。園韓神。松尾。平野。春日。大原野等祭爲小祀。風神祭已上並諸司齋之。鎮花祭已下祭官齋之。俱小祀祭官齋者内裏不齋其道勅使之祭者齋之。

延喜式 卷一 神祇 四時祭上
延喜式 卷一 神祇一 四時祭上
一九

5.2. 構造化テキストとは何か

- 単なるテキストデータではなく、**文章に内在する様々な情報や構造をデータとして反映させたもの**。何らかの**タグ**を伴うことが一般的
- 例：小風尚樹 ← 単なるテキストデータ
 <人名>小風尚樹</人名> ← 構造化テキスト
- これにより、コンピュータがタグ付けされたテキストの意味を解釈できるようになる(コンピュータ可読データと呼ぶ)ことが重要
- **HTML**や**XML**などが有名(MLはマークアップ言語の略)
 - HTML: ウェブブラウザ上で情報を表示する際の言語。タグ数が限られている
 - XML: テキストの意味や構造を表現することに特化。**自由にタグ名を設定できる**

5.3. XMLの限界とText Encoding Initiative

- 例:「人名」という概念を表す自然言語はいくらでも存在する
 - 名前(人の); name of person; 名前; 氏名; ...
 - これらをXMLのタグにしてしまうと、同一概念なのに呼称が揺れてしまい、データの一括取得や比較が難しくなる
- 解決策としてのText Encoding Initiative (TEI)
 - 人文学史資料のテキストを構造化するための国際的枠組み(1987~)
 - 人文学に有用なタグの共通名称・用途を開発し、広く共有することを目的に
 - 世界中の文献学関連プロジェクトで採用されている
 - 競争的資金の成果としてのテキストデータは
今やTEI準拠で発信することが事実上の国際標準
 - 詳しくは、永崎研宣「[歴史データのさまざまな応用](#)」『歴史情報学の教科書』

Cf. 世界各地の人文系プロジェクトでひろく活用

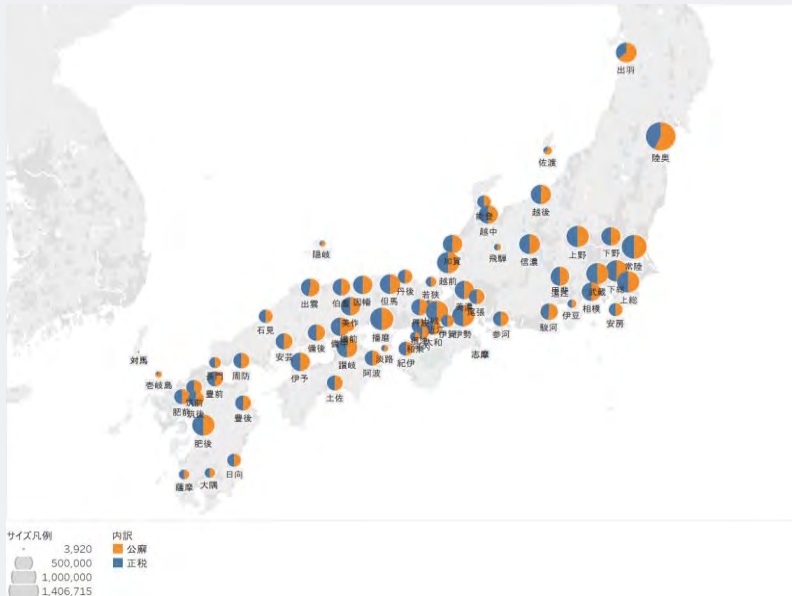
- Perseus Digital Library | ギリシャ・ラテンなどの古典文献学
- Regesta Imperii Online | 中世ドイツ史・ヨーロッパ史の史資料目録
- Shakespeare Digital Texts | Folger図書館やオクスフォード大学など
- ENCCRE | 近世フランス『百科全書』のデジタル学術校訂版
- Van Gogh the Letters | ゴッホの書簡テキストデータベース
- British National Corpus | イギリス英語の1億語のコーパス
- 大蔵経データベース | 台湾や日本など世界各地の仏典プロジェクト
- **延喜式(歴博)** | 律令制下、日本古代の施行細則

...

Cf. 日本におけるTEIの関連動向

- 東アジア／日本語分科会による活動
 - 座長： Charles Muller／永崎研宣
 - 運営委員：岡田一祐／中村覚／永崎研宣
- 東京大学大学院人文社会系研究科「人文情報学研究」での授業展開
- 東京大学附属図書館における公開テキストでの採用
- **国立歴史民俗博物館の延喜式プロジェクト**
- 国文学研究資料館での試行
- 関西大学KU-ORCASプロジェクトでのガイドライン和訳公開
- 大蔵経テキストデータベースでの一部採用
- その他、複数機関での採用予定

6. 延喜式におけるTEIの活用



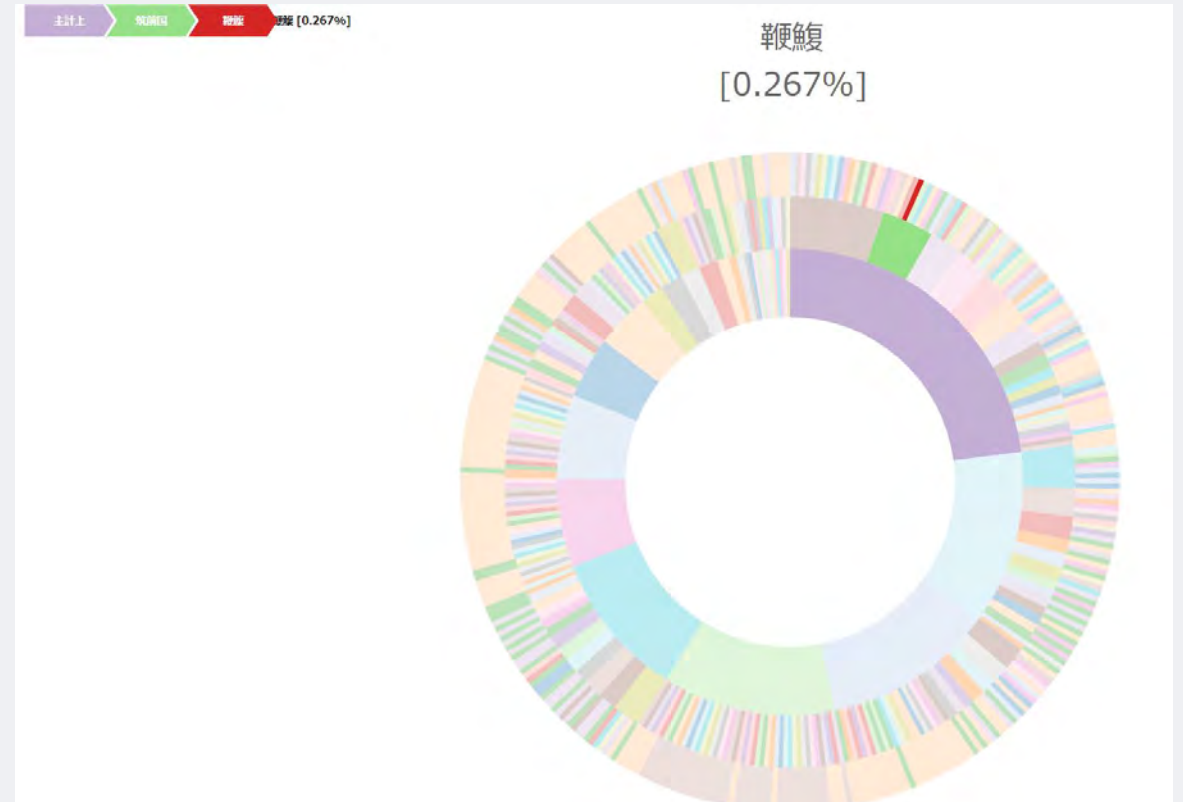
正税・公廩稻額に基づいた各国の一般会計における財政規模の比較 (cf. Tableau Desktopにて可視化)

種別	用途・備考	単価	斤	両	計算式
官司内で消費	齋宮の初齋院官人五位の食法。一日分。月毎支給。初齋院は齋王選出時にのみ臨時。伊勢の齋宮寮から離れた京に所在。別会計か。	2両/1日	88	8	$2\text{両} \times 2\text{人(五位官人人数)} \times 354\text{日} = 1416\text{両}$ (日数は不定。ここでは1年分(30日×6ヶ月(大の月)と29日×6ヶ月(小の月)を合算した354日とする)で計算。換算すると88斤8両)
官司内で消費	齋宮寮の月料(一月分の必要経費)。齋宮寮での消費が予定された執行予算分量。他の項目の執行分を含んで計上されている可能性あり	24斤6両/1月 (13両/1日)	287	10	$24\text{斤}6\text{両} \times 6\text{(大の月)} + [24\text{斤}6\text{両} - (24\text{斤}6\text{両} \div 30)] \times 6\text{(小の月)} = 4602\text{両}(287\text{斤}10\text{両})$
官司内で消費	正月三節の齋宮への供料	3斤/1節	9	0	$3\text{斤} \times 3\text{節}$ ※3節合算で3斤の可能性もあり
官司内で消費	新嘗祭の祭料。齋宮寮より支出		2	10	
収入	安房国の調庸の鰯を送納(年)	300斤/1年	300		

齋宮式に見る東鯨の収支状況の試算

延喜式におけるアワビの呼称一覧

(cf. E2D3にて可視化)



Cf. Naoki Kokaze, 2021 (forthcoming), 'Application of the Methodology for Structuring Historical Financial Records to a Japanese Historical Source along with Financial Information'

7. 課題と抱負

- マークアップ事業には歴史学や文献学、情報学の知識を総動員する必要
- 専門知識とマンパワーをどう確保するか
 - クラウドソーシングなどの手法が採用される場合も (cf. Transcribe Bentham)
 - 関心ある市民、研究者による協力の成果を専門家が質的に検証するなどウェブの力を活用した協働の仕組み作りが求められる場合も
 - DHにおけるプロジェクトデザイン・マネジメント能力の必要性
 - 人材育成・教育における要件設計を誰が担うのか
 - [JADH「人文学における情報リテラシー」研究会](#)などでの活動を展開していきたい

ご清聴ありがとうございました