論文に基づく歴史学知識資源の公開に向けたデータセット記述の精緻化

NLP2026 併設ワークショップ JLR2026
「日本語言語資源の構築と利用性の向上」

亀田 尭宙 (人間文化研究機構)

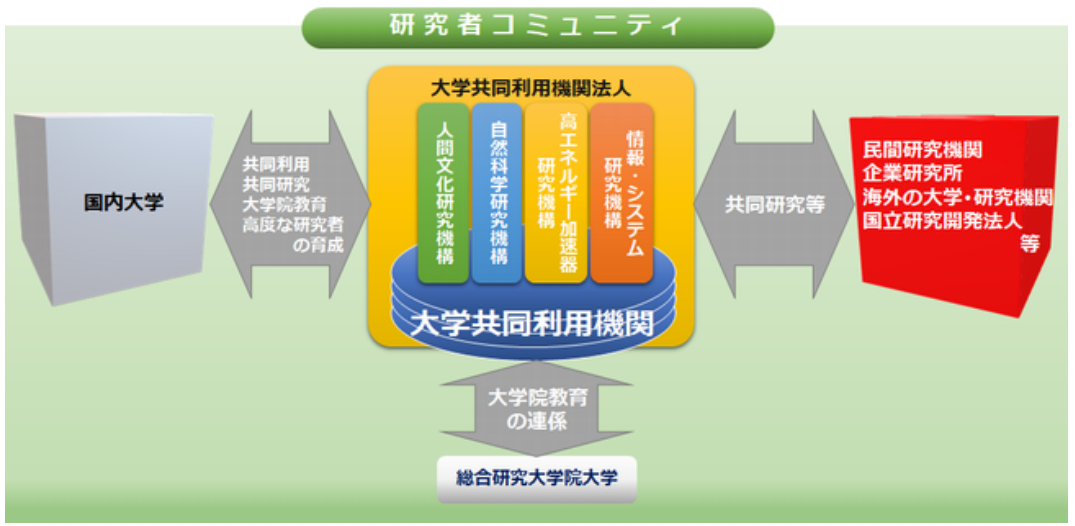
2026年3月13日

人間文化研究機構について

本部に人文情報学（DH）の促進事業があり、私はそこに属してます。

今日は、前職の国立歴史民俗博物館に関わるテーマで「論文に基づく歴史学知識資源」を日本語言語資源の一つとして構築する際の課題と工夫について話したいと思います。

人間文化研究機構について(cont.)



これらの組織は第一義に大学、そして企業などもクライアントとして、「個々の大学では維持することが難しい大規模な実験・観測施設、及びデータベースや研究資料等を整備して、大学等の研究者がこれらを利用し、効果的に先端的な共同研究ができるように」するのがミッションです。

すでにある取り組み

- 国語研コーパスポータル を中心に、研究用コーパスを継続的に公開してきた蓄積がある。
- 歴博の 「みんなで翻刻」テキストデータ と、それをもとにしたNDLの 古典籍OCR学習用データセット のように、市民参加型の翻刻と機械学習用のデータ提供が接続している。
- 国文研は国立情報学研究所大規模言語モデル研究開発センターに古典籍テキストデータの提供する協力を始めている。

今回の取り組み

- 歴史学の論文から史料言及を知識資源として切り出す。
- 各エントリは **史料名 / 史料本文 / 読み（解釈）** と、論文への出典ポインタを持つ。
- 論文本文全体は配布せず、スタンドオフな記述として公開する。
- JLR2026 の呼びかけ文にある「利用性の向上」に寄与する形を目指す。
- この取り組みは、科研費「歴史知識辞書（ナレッジベース）の構築」の支援を受けています。

例: 論文中の該当箇所

これは日光山縁起のこの土地における受容と変容とを示している。「風土記」には巻之十に越後国蒲原郡鹿瀬組実川村農民所蔵として「日光山縁起」が収録されている（1-152 ~ 160）ので編纂にあたっては、もちろんこのことが意識されており、それは割注に示されている。ここで重要なのは、そうした記録に対応する在地の伝承が記載されているということである。

その内容は、赤城山の神、すなわち百足虫を、護摩を修することで鎮めたということから、どちらかというと赤城の神に傾斜した伝承ということができよう。それと同時に空海も登場する点に注意しておきたい。それに加えて大清水池という沼が沼沢村の沼と水脈が通じていて水の増減が共通するので、「夫婦沼」と呼ばれることが記されている。さらに

> 此沼ニ魚アリ、其形チ「ホヤ」ニ類シ口尖レリ、コレヲ安座魚ト云、此沼ト沼沢村ノ沼ノミニ生シテ他ヨリ産スルコトナシ、コレヲ昔ノ八蛇沼ノアトナリト云（5-18）

とされていて、独特の形状の魚が神霊と関連する沼にだけ生息していることが意識されていたことがわかる。これもまた日光山の縁起の変奏の一部と解釈することができるだろう。こうした自然認識が、説話的な伝承のよりどころともなっていたという点に注意しておきたい。伝説化した日光山縁起と地域の自然とが相呼応しながら地域に伝えられていたのである。

(小池 2025年)

例: 抽出される中身

- 史料名: 日光山縁起
- 本文: 此沼ニ魚アリ、其形チ「ホヤ」ニ類シ口尖レリ、コレヲ安座魚ト云、此沼ト沼沢村ノ沼ノミニ生シテ他ヨリ産スルコトナシ、コレヲ昔ノ八蛇沼ノアトナリト云(5–18)
- 読み: 独特の形状の魚が神霊と関連する沼だけに生息していることが意識されていた, 日光山縁起と地域の自然が相呼応しながら伝えられていた など
- 出典ポインタ: 小池淳一「会津における歴史文化研究拠点の伝承と記録」(小池 2025年)

なぜ論文に基づく知識資源が必要か

- 資料のデジタル化だけでは、それをどう読むべきかがわからない。
- すでに論文の中では専門家によって同定・読み・異説整理がなされているので、それをデータ化したい。
- 日本語は学習用資源が多い方ではあるが、文化的・地域的知識に踏み込んだ評価資源はまだ薄い(Romanou ほか 2024年)。

論文のテキストデータがあればLLMがよしなにやってくれるのでは？

- 出版社にお金を払って再配布も含めて使えるテキストデータを整備するのも検討中。
- ただ、まずは少量でも質の高いデータで、文化的・地域的知識に関して何が有効な資源になるかを検証したい。
- また、テキストが手に入っても、著者の権利への配慮や、利用者がその「読み（解釈）」の正当性を出典まで遡って確認できることは別途必要になる。
- そこで今回は、フルテキスト整備の代わりではなく、その先でも必要な **史料名 / 読み（解釈） / 注意点 / 出典ポインタ** の層を先に設計する。

全部を配らなくても公開できる層

- 既存の権利保有コンテンツと競合しない粒度で、構造化データを公開する先例がある。
- 人文学オープンデータ共同利用センターの構築している地名辞書では、ジャパンナレッジ項目ID、地名、読み、緯度経度などを１レコードとするデータセットを公開し、詳細情報はジャパンナレッジ経由で『日本歴史地名大系』へ導くことで出版社と合意できている。
- 今回も同様に、論文本文そのものではなく、**史料名 / 読み（解釈） / 出典ポインタ** の層をまず公開対象として、論文本体は（無料だが）別プラットフォームに誘導する形式になっている。

対象にした論文誌

JSTで（エンバーゴ付き）オープンアクセスになっているものと、紀要。

- 『日本史研究』 日本史研究会
- 『史学雑誌』 公益財団法人史学会
- 『国立歴史民俗博物館研究報告』 国立歴史民俗博物館

> 『日本史研究』2026年2月号 では「会財政の現状について」が掲載され、二年連続の大幅な赤字への危機感が共有されています。私たちは、こうした状況も踏まえつつ、雑誌に対価を払いながらコーパスを作る方向で、まずはパイロット的な試行錯誤をしています。

オープンライセンスにならない雑誌

- PDF 解析のツール GROBID の日本語論文適用を狙う先行研究でも、データセット公開を見据えて CC BY / CC BY-SA 論文だけを対象にしている (嘉本 ほか 2025年)。つまり、オープンライセンスでない論文は、データセットとして再配布・共有する素材には一般にしにくい。
- ところが日本史分野では、自由な二次利用を積極的に啓蒙する必要はなく、ライセンス付与を急いで議論する必要もない、という認識が示されている (JPCOARコンテンツ流通促進作業部会 DOIチーム 2026年)。本に再編することがあるような分野で、著者側にとってもメリットがあるはずなのに...？

PD BY の問題

> パブリックドメインである古文書の利用に出典の明記を求めることは、法的な根拠がないことから、利用の条件としてではなく、あくまで「お願い」として記載することも考えられます

(DH権利問題支援ツール検討会 2024年)

- CC BY-SA の Wikipedia から作った、日本:JPYといった ISO3166 コードの csvデータ（著作権が無いと推定される）が CC BY で出されている、という事例が実際にある。「著作権が無いと推定」し上書きしてもそれが誤っている可能性があり、逆に「元のママ」CC BY にしてもPD BY 状況になる。
- contributions と著作権上の条件を分けて書き、あわせて判断根拠を残すのが理想。

来歴の重要性（1）

- データロンダリング / ライセンスロンダリングは、出典や条件をぼかしたまま、より正当で緩いデータとして流通させる問題。
- 実際にファインチューニング用データセット監査で、70%超のライセンス欠落、50%超のライセンス誤りで緩い条件に変えられていた例が多かった (Longpre ほか 2024年)。
- したがって、出典、作成者、ライセンス、派生関係を一緒に運ぶ来歴 (provenance) メタデータは、少なくともロンダリングに気づくための前提になる。改ざん検知可能などの発展も (Coalition for Content Provenance and Authenticity 2025年)。

来歴の重要性（2）

- そもそも学術的に、来歴が不明なデータは、そのままでは根拠として使いにくい（cf. 校訂者を気にしたりする）。
- したがって来歴つきのデータ流通は、ライセンスロンダリング対策であると同時に、データの信頼性の確保、引用可能性や批判的検討可能性を支える研究基盤でもある。
- だから今回のデータセットでは、出典ポインタ、抽出判断、関与者の情報を持たせたい。

> 校訂には基本的に著作権がないものの、ある場合も、著作隣接権があるとされる場合もある (石岡 2009年)。

All contributors

- All Contributors は、`README` や `CONTRIBUTORS.md` にコード以外の貢献も含めて機械可読な形で並べるための取り組み。
  - プロジェクトへの参加を歓迎するための枠組みでもある。
- 今回の文脈では、著者・抽出者・校正者・画像提供者・権利確認者といった役割を記述したい。
- こうしておくと、著作権ライセンスだけに無理に責任や謝辞を押し込まずに、誰が何に関わったのかを示せる。

> 注: AI学習が著作権法30条の4の権利制限に入る限り、`CC BY` の `BY` 条件を履行する法的必要はない

他の well-known files

- NOTICE.md 権利留保、再利用時の注意、出典表記のお願い、第三者資料の除外などをまとめる。
- ETHICS.md 想定する利用、避けてほしい利用、データの限界、利用時の確認事項を分けて書く。Deon の examples のような雛形が参考になる。

→NOTICE.mdやETHICS.mdに書かれた項目を守った利用ができているかのチェックする際のプロンプトにも使える。

まとめ

- 歴史学の論文そのままデータ化ではなく、`史料名 / 史料本文 / 読み（解釈）` と出典ポインタを切り出して出す。
- その際の課題は、オープンライセンスではない論文誌、`PD` `BY` のような実務上の混線、ライセンスロンダリングを防ぐための来歴管理。
- だから、本文の再配布可否だけでなく、出典、判断根拠、関与者、利用上の注意を分けて持つ設計が必要になる。
- `CONTRIBUTORS.md`、`NOTICE.md`、`ETHICS.md` なども含めて公開単位を設計することで、研究利用にもAI利用にも耐える資源に近づける。
- 人文系の研究者個人レベルでのデータ公開を加速したい。

LICENSE

- このスライド群のうち、著者が今回書いた本文と、今回自作した図表は CC BY 4.0 で公開する。
- ただし、第三者由来の素材はその範囲に含めない。具体的には、人間文化研究機構の図、論文の紙面画像、論文からの引用などである。
- 詳細な除外対象は NOTICE.md にまとめてある。

References

- Coalition for Content Provenance and Authenticity. 2025年. *C2PA Specification Version 2.2*. https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html.
- DH権利問題支援ツール検討会. 2024年. *デジタル・ヒューマニティーズ（DH）研究に関する権利問題ガイド*. NIHU DH 人間文化研究機構DH促進事業ウェブサイト. https://dh.nihu.jp/projects/right/post/kenri_001.
- JPCOARコンテンツ流通促進作業部会 DOIチーム. 2026年. *DOI・ライセンスに関するヒアリング調査について（2022年度版）*. JPCOARリポジトリ; オープンアクセスリポジトリ推進協会. https://doi.org/10.34477/0002000689.
- Longpre, Shayne, Robert Mahari, Angela Chen, ほか. 2024年. 「A Large-Scale Audit of Dataset Licensing and Attribution in AI」. *Nature Machine Intelligence* 6 (12): 1419–31. https://doi.org/10.1038/s42256-024-00878-8.
- Romanou, Angelika, Negar Foroutan, Anna Sotnikova, ほか. 2024年. *INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge*. https://arxiv.org/abs/2411.19799.
- 嘉本名晋, 梅澤悠河, 長尾浩良, 桂井麻里衣. 2025年. 「日本語論文に特化したPDF文書解析器の構築と性能評価」. *言語処理学会第31回年次大会 発表論文集*, 3月. https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/A3-4.pdf.

- 小池淳一. 2025年. 「[論文] 会津における歴史文化研究拠点の伝承と記録 : 『新編会津風土記』の分析」. *国立歴史民俗博物館研究報告*, no. 264 (2月): 35–60. https://doi.org/10.24619/0002000213.
- 石岡克俊. 2009年. *「校訂」の著作権法における位置*. KEO discussion paper 116. Keio Economic Observatory. https://doi.org/10.14991/004.00000116-0001.