

第8回東京大学学術資産アーカイブ化推進室主催セミナー 2025／01／28

光源氏はデジタルアーカイブの夢を見るか？

人文社会科学における 研究資料デジタル化の現在とこれから

デジタル化は人々の「欲望」を満たし続けるか？

東京大学デジタルアーカイブ化推進室は5年目に入り、学内にある学術資産のデジタル公開は着実に進んでいます。一方でデジタルアーカイブの構築・運用に関する課題は担当室によって際々にあります。また、デジタルアーカイブ関連の技術は日進月歩で、最新情報をキャッチアップすることはやがていっそう難しくなっていきます。

そこで本学の教員と附属図書館職員有志で運用している「デジタル源氏物語」を題材に、その取組から見てきた課題や、「デジタル源氏物語」で利用されている技術の現状を共有することで、今後のデジタルアーカイブ構築・運用に役立つ情報を共有するためのセミナーを開催します。

「源氏物語」という古典作品のデジタル活用から見えてきた様々なトピックスをもとに、デジタルアーカイブの今後について一緒に考えてみませんか？

第8回東京大学学術資産アーカイブ化推進室主催セミナー
光源氏は
デジタルアーカイブの
夢を見るか？

参加申込み
下記URLからお申し込みください
(要予約・定員50名を超過しません)
<https://www.lib.u-tokyo.ac.jp/lib/contents/event/20250128/>
参加費無料

プログラム ※プログラムのタイトルは全て仮題

セミナー趣旨説明
田村隆
東京大学大学院総合文化研究科 准教授
ミニマムなデジタルアーカイブシステム構築の実践について
阿部悠司
東京大学大学院人文社会系研究科 文化資源学専攻 准教授
デジタルアーカイブに関する広報活動について
徳元美穂子
経済大学附属図書館 情報サービス課 准教授
OCR技術の最新動向について
高田直子
国立国会図書館 学術情報部 学術情報課 主任
TEI(Text Encoding Initiative)の概要と活用について
鈴置伸子
東京大学情報システム部 情報基盤課 学術情報チームデジタル・ライブラリ担当
編纂部 東京大学経済学図書館 (京大蔵書館) 准教授
デジタルアーカイブにおけるAI活用の最新動向について
大隅一博
東京大学大学院人文社会系研究科 准教授
人文社会科学分野における研究資料デジタル化の
現在とこれから
山本和明
人間文化研究機構 国文学研究資料館 研究主幹・教授
研究主幹・教授 (併) 基盤データセンター長

【お問い合わせ】
東京大学学術資産アーカイブ化推進室
お問合わせ先: lib-arc@lib.u-tokyo.ac.jp
03-5841-2607
※開催のため、参加に配慮が必要な方は
事前にご連絡ください。

主催: 東京大学学術資産アーカイブ化推進室 U-Tokyo 50 国文学研究資料館

人間文化研究機構
国文学研究資料館 研究主幹・教授
(併)基盤データセンター長

山本 和明

自己紹介:山本和明(YAMAMOTO Kazuaki)



大阪の私立大学教授等を経て
2013年10月～現在まで
国文研で歴史的典籍NW事業等に従事。

現在、国文研DDHプロジェクトで
基盤データセンター長として、
画像・テキスト作成事業を推進

研究分野:19世紀日本文学

DB利用の**エンド・ユーザー**



日本語の歴史的典籍の国際共同研究ネットワーク構築計画(略称:歴史的典籍NW事業)

- ・文部科学省の大規模学術フロンティア促進事業に**人文・社会系ではじめて採択**
- ・国内外の大学等の研究機関・図書館等と連携し、**歴史的典籍30万点のデジタル画像化**
- ・諸分野の研究利用に資するべく「**国書データベース**」として**Web上で公開**
- ・その画像を用いた**国際的な共同研究のネットワークを構築**

(事業期間:2014年度~2023年度)。

データ駆動による課題解決型人文学の創成(略称:国文研DDHプロジェクト)

- ・**大量の古典籍の電子化・テキスト化**による一般への開放、及びデータの構造化・体系化と利活用
- ・古典を開く **現代の課題から古典の知を発見**する
- ・研究基盤の維持・発展が不可欠
- ・AIによる機械可読データの生成に向けて

(事業期間:2024年度~2033年度)

データの
集積

データインフラ
ストラクチャの構築

古典籍

- ・**15万点の追加画像**
- ・**テキストデータ大幅増**
- ・**メタデータの標準化** 等

国書データベース

書誌検索

著作検索

著者検索

画像タグ検索

全文検索

源氏物語画帖

光源氏



いつでも

どこでも

誰でも

画像の公開:学問の進化へ

デジタル源氏物語をネタ(酒の肴)にして…

◆ミニマムなデジタルアーカイブシステム構築の実践について(阿達藍留氏)
【システム】 いかに簡便につくれるか？ システム例

◆デジタルアーカイブに関する広報活動について(徳元美智子氏)
【広報】 いかに人々に知らしめるか 広報戦略

◆OCR技術の最新動向について(青池亨氏)
【テキスト作成】 いかにデジタル画像からテキストを増やすか

◆TEI(Text Encoding Initiative)の概要と活用について(鈴置怜子氏・福嶋彩乃氏)
【テキスト作成】 テキストの構造化・高度化

◆デジタルアーカイブにおけるAI活用の最新動向について(大向一輝氏)
【最新動向】 AIによる可能性とその動向

→普及させること、チャレンジしてもらうこと

【良い面】

1. 多様な資料へのアクセス

- ・複数の機関が所蔵する異なるバージョンや挿絵、関連資料がデジタル化され、オンラインで閲覧可能です。

2. 研究支援機能の充実

- ・原文画像とテキストを並べて表示する機能や、異なるバージョン間の比較機能、現代語訳の参照など、研究者や学生が詳細な分析を行うためのツールが提供されています。

3. 教育・学習の促進

- ・くずし字の認識サービスや、挿絵の比較表示、源氏百人一首の顔比較など、視覚的・直感的に理解を深めるための機能があり、教育現場や一般の学習者にも有用です。

4. 国際的なアクセスと普及

- ・インターネットを通じて世界中からアクセス可能であり国際的な研究や普及に貢献。

【悪い面】

1. 技術的な制約と課題

- ・画像の解像度や表示速度、ユーザーインターフェースの使いやすさなど、改善の余地がある場合があります。

2. 原本の物理的特性の喪失

- ・デジタル化により、原本の質感や紙質、筆跡の微妙なニュアンスなど、物理的な特性を直接感じられません。

3. デジタルデータの保存と永続性

- ・技術の進展やフォーマットの変化により、長期的な保存やアクセスに課題が生じる可能性があります。

4. 利用者のデジタルリテラシーが必要

- ・利用者側にも一定のデジタルリテラシーが求められ、全ての人が容易に利用できるとは限りません。

デジタル源氏：Grokに聴いてみた！

【良い面】

アクセス性の向上

『源氏物語』写本を誰でも容易に閲覧可能。
一般の人々も日本の古典文学に触れる機会が増。

比較研究の促進

複数の写本を比較することが容易に。

保存と保護

物理的な劣化を防ぎ、文化財の長期的な保存と保護に貢献。

新しい研究手法の導入

くずし字OCR(光学文字認識)の活用により、検索機能が向上し、特定のシーンや語句を効率的に探すことができ、研究の効率化に寄与。

【悪い面】

デジタル化の限界

写本の質感や微細な筆跡のニュアンスなど。これにより、研究者が得られる情報に限界が生じることもある。

技術的な問題

デジタル化プロジェクトの維持には、継続的な技術的サポートと更新が必要。

著作権やアクセシビリティ

資料がどの程度公開されるか、あるいは誰でもアクセスできるかは、著作権やデータの所有権に関する問題が浮上。

教育的価値の変化

デジタル資料への移行は、学問的価値や学習者の学習体験に影響を与える可能性がある。例えば、紙の質感やページをめくる感覚などが失われ、古典文学の体験が一部変わるかも。

デジタル源氏を支える技術とひと

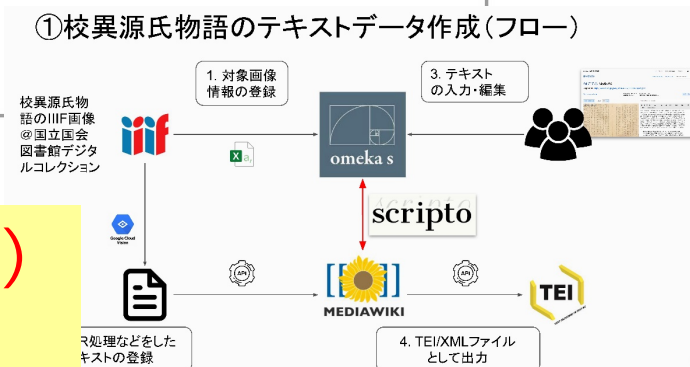
中村覚「デジタル源氏物語」を支える技術:IIIFとTEIより
(2020年11月7日第6回日本語の歴史的典籍国際研究集会発表資料)



	方法	作成したデータ
①校異源氏物語のテキストデータ作成	- Omeka S - Scripto	TEI準拠のXMLファイル
②公開画像への校異源氏物語の頁数付与	くずし字OCR	くずし字OCRテキスト
	- 編集距離 - 人手による確認	IIIF Curation API準拠のJSONファイル
③校異源氏物語と現代語訳の対応づけ	- 青空文庫テキスト <anchor/>タグ	TEI準拠のXMLファイル

省力化に
貢献

構築したデータ(主な作業員5名!!)
しかもボランティアですよ?



不易と流行

【コンテンツ】

画像 テキスト

- ・ある程度決まった形式で蓄積しさえすれば、その価値は減らない
- ・つくること、データの維持管理が大変
- ・機械的処理の工夫・管理の工夫が今後の課題

コンテンツは死なない

【 ツール 】

ビューワ **ツール**

- ・ツールは日進月歩(百花繚乱)
- ・OSの更新などの影響を受ける(技術的サポート)
- ・脚光を浴びやすい
(言い換えれば陳腐化も早い)
- ・尖った技術か、枯れた技術か

最新技術のデパートか、
ユーザーを意識した造りか

ツール： その思想＝「志」は那邊に？



画面遷移
TOP



◆エンドユーザーとしての私感
(直感的に・無いものねだりだけど)

- ・誰のためのツールか？
想定する人文研究者像は
- ・どういう場面で使うのか？
大学の演習？講義？

・利用者のデジタルリテラシーをどう考えているか ←←日本文学研究者 もっとも疎い存在？(あと10年はかかる)

→→利用者が、主体的につくったという「仕掛け」が欲しい



帖ごとでの構成

ツール： 今後の展望は？

今後の展望(中村覚氏の資料より)

● 第三者・各所蔵機関が追加可能な仕組みの構築

- ・IIIF画像の準備
- ・くずし字OCRの実施
- ・ページ数の推論結果に対する人手によるチェック等々

● 対象

源氏物語／その他の古典籍

- ・維持することが大変（枯れた技術か尖った技術か？）
- ・源氏のみでいいのか？
 - ←←源氏は使えるデータが存在したという点
- ・ツールの利用方法を如何に教えるか
 - 海外：研究者と司書の関わり大　日本は？
- ・ボランティアのモチベーション維持を（賞を取った今だから）

いつまでも「ウラ」ではなく(笑) … 人肌脱いでみませんか？

IIIFの例 

【歴史】

画像APIは2011年末に、大英図書館とスタンフォード大学が、アンドリュー・W・メロン財団からの資金援助を受けて、世界有数の研究図書館（ボドリアン図書館〔オックスフォード大学〕、フランス国立図書館、ノルウェー国立図書館、ロスアラモス国立図書館、コーネル大学）とともに、画像配信のための相互運用可能なフレームワークを共同で構築する1年間の取り組みに着手し、2012年にバージョン1.0が公開され、プレゼンテーションAPIのバージョン1.0が2013年に、2016年に検索APIが公開された。画像ベースのリソースへの均一でリッチなアクセスを前例のないレベルで学者に提供されている。

<https://www.cni.org/topics/information-access-retrieval/international-image-interoperability-framework>

https://ja.wikipedia.org/wiki/International_Image_Interoperability_Framework

束となってかかる 源氏から古典籍用ツールへ(汎用性)
国文研も応援(協力)します

「世界」にうってでる 広報とセッション



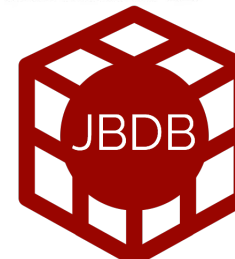
昨年の大河とリンクした
SNSは頭が下がる思い
です



対面で聴くメリット

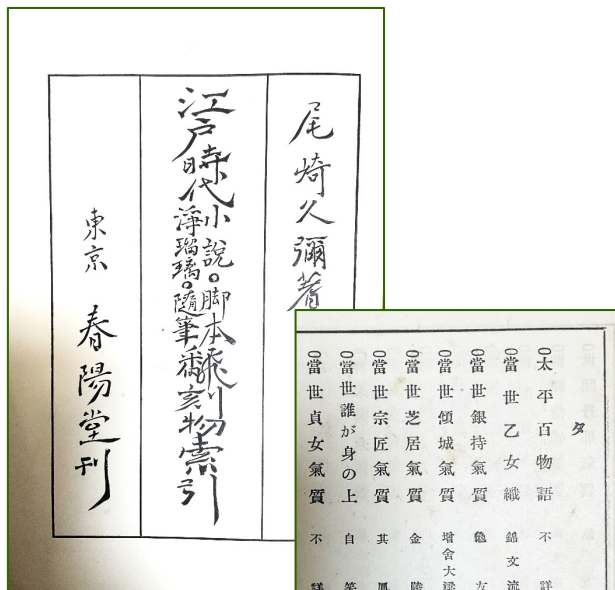


eajrs
European Association of
Japanese Resource Specialists 日本資料専門家欧州協会



日本の人名DB デジタル文学地図

コンテンツ:テキストが足りないって本当？



国文学・アーカイブズ学論文データベース

国文学研究資料館 > 電子資料館 > 国文学・アーカイブズ学論文データベース > 国文学論文

件数: 635479 件

TOPへ

カテゴリ

時代分類/分野

検索結果一覧

検索条件: 全ての項目 <翻>

検索結果: 7187件中 1 - 50 を表示

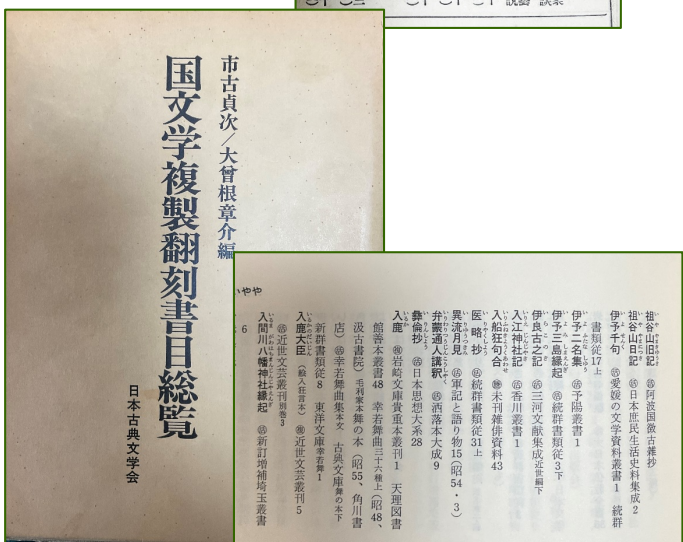
50件ずつ表示 IDでソート 表示

No.	題名	執筆者名	誌名	巻号	通巻	発表年	国文研請求記号	時代分類	分野	英文題	英文誌
1	<翻>資料翻刻『神三郎神社参詣記』(二)	桜井治男	皇学館大学神道研究所紀要	2		1986	J00044	国文学一般	古典文学		
2	<翻>東大寺図書館蔵仏名会法則二種	関口静雄	宇部国文研究	17		1986	J00050	国文学一般	古典文学		
3	<翻>古筆切目安 (副子・索引・解題)	高田信敬	鶴見女子大学紀要	23		1986	J00080	国文学一般	古典文学		
4	<翻>山内神社宝物資料館蔵『古筆手鑑』(甲)について<前承>	徳満澄雄	高知女子大学紀要	34		1986	J00110	国文学一般	古典文学		

国文研 国文学・アーカイブズ学論文DB
 <翻>で検索 → 7187件 国文学関係

テキストはある！
 利用する仕組みがないだけ

昔から活字化がなされてきた人文学
 商業ベース／大学等の紀要等々
 復刊の扱い
 翻字ではなく、**釈文**という形が多い



テキスト いかにデータを集約するか これが難題

【現状】

- ・利用したい機関・人間が許諾を求めなくてはいけない(許諾手続きという壁)
- ・大量にこなす(紐付ける)ための員数＝予算の必要(手作業)
- ・データではなく誌面として存在→OCRが必要(国会図書館ではウラにデータ持っている場合もあり)



いくらテキストがあってもそれを活かす
(使える)工夫がなければはじまらない

オープンな
利用を推進

図書館

紀要等の公開データを
遡及してオープン化
宣言

各研究者

自身の研究成果の
オープン意思表示
(リサーチマップ等)

学会

学会構成員の
遵守事項にする
データの集約に寄与

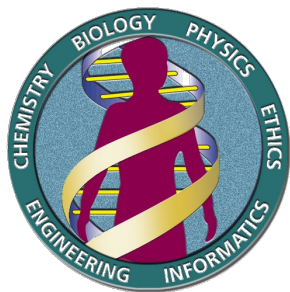
→こういうことを言うと国文研やれという「声」続々…

なぜ出来ないと言いつけるのですか？ ～すれば出来るという発想へ

Human Genome Project

このプロジェクトは1990年にアメリカのエネルギー省(DOE)と国立衛生研究所によって30億ドルの予算が組まれて発足し、15年間での完了が計画されていた。発足後、プロジェクトはアメリカ、イギリス、フランス、ドイツ、中国、日本の国際的協力の拡大と、ゲノム科学の進歩(特に配列解析技術)、およびコンピュータ関連技術の大幅な進歩により、2003年に完了した。

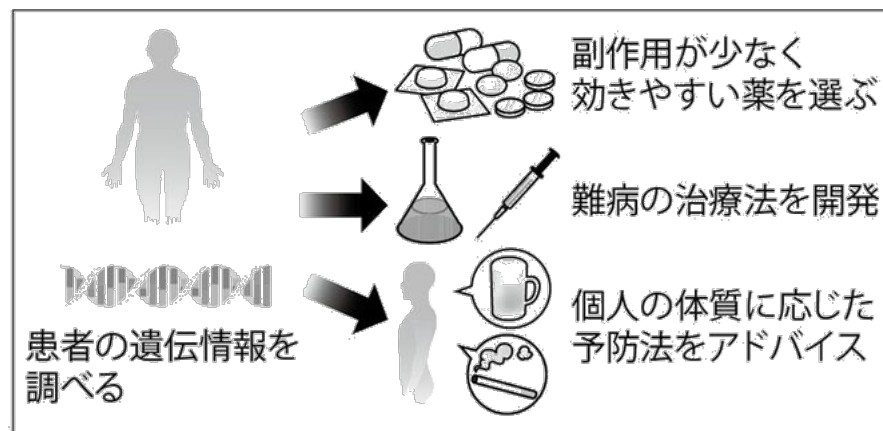
<https://doe-humangenomeproject.ornl.gov/>



- ・人々からのDNAサンプルの提供
- ・国際化
- ・コンピュータの進歩等々

<https://www.kanehisa.jp/documents/brochure/japanese/QA.html#Q8>

■ 遺伝情報を活用した「ゲノム医療」のイメージ





もっとも考えなくてはならない懸案が、「デジタル保存」である。「デジタル化(digitization)」と「デジタル保存(digital preservation)」は全くの別物であり、デジタル化を終えてのち、継続して保存、公開を維持して行くのが実は一番難しい。

継続的な保存および配信を考えてこそそのDAでなくてはならず、また発見に繋がるよう他機関との連携も見据えなければならない。文字通りのWEB(蜘蛛の巣)化を果たしたDBとなり得るかがカギとなる。

既存の人文学の研究者も、与えられることに慣れすぎていないか。DAもそうだが、30万点を公開しても、国文研がするのは当たり前と考えている節すらある。

大規模に予算を獲得し、事業として進めることに甚だ無関心だ。すこし厳しいかもしれないが、誰が作るのか、誰が維持するのか(予算獲得を含め)。いつまでも他人任せで良いのかと申しあげたいのである。束となっていく姿勢なしには、人文学の未来はない。

デジタル化の弊害 物に触れる機会の喪失 すでに危険区域に突入

(一例)

- ・付け爪をして和本にふれる
- ・和本を前にして他の人としゃべり出す
- ・手にとって和本を読もうとする
- ・扱える図書館司書の知識と人数

等々

(再掲)**教育的価値の変化**

デジタル資料への移行は、学問的価値や学習者の学習体験に影響を与える可能性がある。例えば、紙の質感やページをめくる感覚などが失われ、古典文学の体験が一部変わるかも。