様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０６２６

研究課題名（和文）IIIFとTEIを用いたオンライン翻刻支援システムの開発

研究課題名（英文）Development of Online Transcription System with IIIF and TEI

研究代表者
中村　覚（Nakamura, Satoru）

東京大学・史料編纂所・助教

研究者番号：８０８０２７４３

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究の目的は、オンライン上で複数の利用者が協力して史料を翻刻可能なシステムを開発することである。特に、画像共有のための国際標準であるIIIFや、人文学資料向けの構造化ルールを定めるTEI等の国際規格に適合させることで、幅広い用途・国際的に活用可能なシステムを構築した。本システムを用いて、画像と多様なテキストデータを関連付け、「源氏物語」の本文研究支援を目指すウェブサイト「デジタル源氏物語」を公開した。更に、くずし字OCRと編集距離を活用し、テキストデータが類似する写本・版本の画像を自動推薦する機能を提供する「デジタル源氏物語（AI画像検索版）」の公開も行った。

研究成果の学術的意義や社会的意義
本研究はIIIFおよびTEIなどの国際規格に準拠した、オンライン上での史料翻刻システムの開発により、人文情報学分野の発展に寄与した。また多様なテキストデータと画像を統合することで、「デジタル源氏物語」ウェブサイトの公開など、学術研究基盤の強化に貢献した。さらにくずし字OCRの利用、およびテキストデータが類似する写本・版本の画像を自動的に推奨する機能の組み合わせにより、歴史資料の新しい活用方法を提案した。デジタルアーカイブ学会と情報処理学会から評価されたこれらの取り組みは、歴史資料へのアクセスを容易とし、国際的な学術研究を促進する。

研究成果の概要（英文）：The objective of this research is to construct a system that allows multiple users to collaboratively transcribe historical materials online. Significantly, by ensuring compliance with international standards such as IIIF, an international standard for image sharing, and TEI, which stipulates structured rules for humanities resources, we have developed a versatile and internationally applicable system. Utilizing this system, we have associated diverse textual data with images about “The Tale of Genji” and launched the “Digital Tale of Genji” website, aiming to facilitate scholarly research on the text. Furthermore, we have released the “Digital Tale of Genji (AI Image Search)”, which provides features that utilize Kuzushiji OCR and edit distance to automatically recommend images of manuscripts and editions with similar textual data.

研究分野：人文情報学

キーワード：IIIF　TEI　RDF

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　C－１９，F－１９－１，Z－１９（共通）

１．研究開始当初の背景

今日多くの機関から古典籍の写本・版本のデジタル画像が公開されている．また IIIF（International Image Interoperability Framework，画像共有のための国際規格）に対応した画像公開も積極的に進められている．例えば源氏物語については，国立国会図書館，国文学研究資料館，京都大学，九州大学，東京大学，米国議会図書館などで IIIF 画像が公開されている．しかし，これらの諸本からある一つの場面を確認したいとき，テキスト検索ができないことが一般的であり，多数の画像から目当ての場面を探し出すにはコストがかかる．源氏物語の場合，例えば国文学研究資料館では全巻（桐壺巻から夢浮橋巻まで）を一つのアイテム（IIIF マニフェスト）にまとめて公開しており，約 2,000 枚の画像から構成される．一方，巻毎にアイテム（IIIF マニフェスト）を公開している機関もあるが，この場合，平均約 40 枚の画像から各巻が構成される．これらの画像から，例えば「夕顔」という文字列が含まれる「校異源氏物語」の 743 頁の場面は，「九州大学 所蔵 源氏物語 古活字版」では 22 巻の（80 枚中の）52 枚目，「東京大学文学部国文学研究室所蔵本（国文学研究資料館提供）」では（2328 枚中の）866 枚目が該当する．諸本の比較を支援する環境が求められる．

２．研究の目的

本研究の目的は，オンライン上で複数の利用者が共同して史料を翻刻可能なシステムを開発することである．特に IIIF や，人文学資料向けの構造化ルールを定める TEI（Text Encoding Initiative）等の国際規格に適合させることで，幅広い用途・国際的に活用可能なシステム構築を目指す．これにより，諸本の比較を支援するシステム開発などにつなげる．特に，『源氏物語』に関する様々な関連データを収集・作成し，それらを結びつけることで，『源氏物語』研究に加え，古典籍を利用した教育・研究活動の一助となる環境の提案を目的としたシステムである「デジタル源氏物語[1]」への適用を通じて，構築したシステムの有用性を検証する．

デジタル源氏物語が提供する機能の例を図 1 に示す．画面右上部には校異源氏物語のテキストデータを表示し，画面右下部には青空文庫で公開されている与謝野晶子による現代語訳[2]を表示している．これらのテキスト間で対応づけがなされている場合には，各々のテキストをクリックすることで，もう一方のテキストの対応箇所がハイライト表示される．また，画面右上部のテキストについて，頁毎に IIIF アイコンが表示される．このアイコンをクリックすることで，国立国会図書館，東京大学，九州大学等で公開されている画像が画面左部のビューア上で表示される．この時，利用者が選択した校異源氏物語の頁数に該当する画像箇所がフォーカスされる．

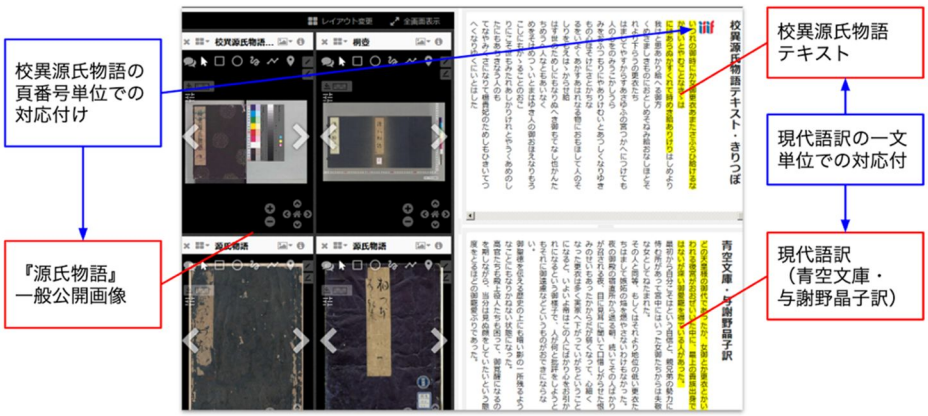

図 1 デジタル源氏物語の提供機能例

３．研究の方法

2 で述べた「デジタル源氏物語」が提供する機能の実現にあたっては，以下のデータが必要となる．以下，それぞれのデータ作成方法，それを支援するシステムについて述べる．

- 校異源氏物語のテキストデータ作成
- 公開画像への校異源氏物語の頁数付与
- 校異源氏物語と現代語訳の対応付け

3.1 校異源氏物語のテキストデータ作成

国立国会図書館デジタルコレクションで公開されている校異源氏物語画像を参照しつつ，テキスト化を行う．作業の効率化のため，事前に Google Cloud Vision API を用いた OCR 処理などを施し，その結果を修正する作業とした．この作業には，Omeka S のプラグインとして構築されている Scripto を使用した．Scripto は MediaWiki を併用し，Omeka に登録済みの画像に対して，翻刻機能を提供するものである．図 2 にその画面例を示す．

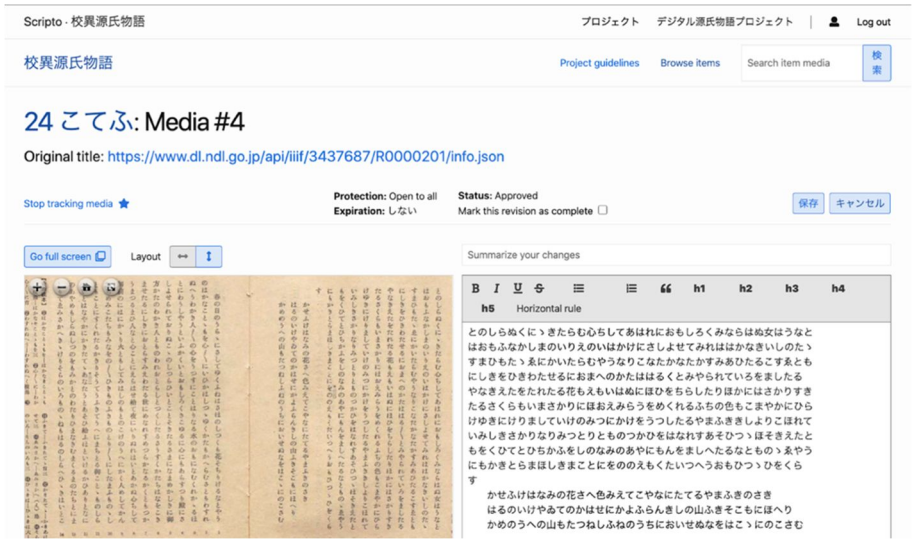

図 2 Scripto を用いた翻刻画面の例

画面左部に画像が表示され，画面右部にテキストエディタが表示される．この画面を使用することで，画像を閲覧しながら，テキストデータの作成を行う．そのほか，複数プロジェクトの作成や，プロジェクト毎のガイドライン（編集方針など）の作成，Reviewer（査読者）の設定などが可能である．編集したテキストデータは MediaWiki に格納されるため，MediaWiki API を使用して，データの取得，TEI/XML 形式への変換などを行う．

3.2 公開画像への校異源氏物語の頁数付与

公開画像への校異源氏物語の頁数付与については，まず CODH（Center for Open Data in the Humanities）が開発している「KuroNet くずし字認識サービス[3]」を利用し，対象資料（例：東大本源氏物語）の全コマに OCR 処理を実施した．これを利用することで，図 3 に示すように，行単位のテキストデータを生成することができる．なお，本作業に含まれる画像の切り取り，切り取り画像の登録，くずし字 OCR の実行，自動テキスト化処理の実行については，Selenium を用いて自動的に行った．

図 3 校異源氏物語の頁数の自動付与

次に，くずし OCR によって作成したテキストデータと，作成した校異源氏物語テキストの各頁の先頭行について，編集距離を算出した．そして，類似度が最も高い行に対して，校異源氏物語の頁数を仮に付与し，この結果を人手で確認する体制をとった．これにより，くずし字を含む画像のみを使って校異源氏物語の頁数を付与していく作業に比べて，専門家の作業の効率化と，くずし字を読むことができない作業者の参画も可能となった．なお，頁数の自動付与の結果は，巻によって精度にばらつきが見られたが，専門家が事前に人手で付与した結果と比較して，概ね 90%

程度の精度（F 値）で正しく推定することができた．

3.3 校異源氏物語と現代語訳の対応づけ

校異源氏物語と現代語訳の対応づけについては，まず青空文庫で公開されている与謝野晶子現代語訳の HTML ファイルから，TEI/XML ファイルを作成した． 次に，作成した校異源氏物語テキストデータに対して，現代語訳の文 ID を<anchor/>タグを使用して挿入した．この作業にあたっては，図 4 に示す，本作業を支援するウェブアプリケーションを作成した．本アプリケーションでは，画面左部に Google ドキュメントを表示し，画面右部には TEI/XML ファイルを表示する．校異源氏物語のテキストデータを画面左部に，現代語訳の TEI/XML ファイルを画面右部で表示することで，Google ドキュメントを使用して，複数人が共同で現代語訳の文 ID を挿入する環境を構築した．なお，画面右部の現代語訳の文 ID をワンクリックでコピー可能な機能などを提供し，ID の挿入作業を効率化する工夫を施している．この作業結果について，Google Docs API を使用して，ID が付与されたテキストデータを取得し，TEI/XML 形式に変換して保存した．

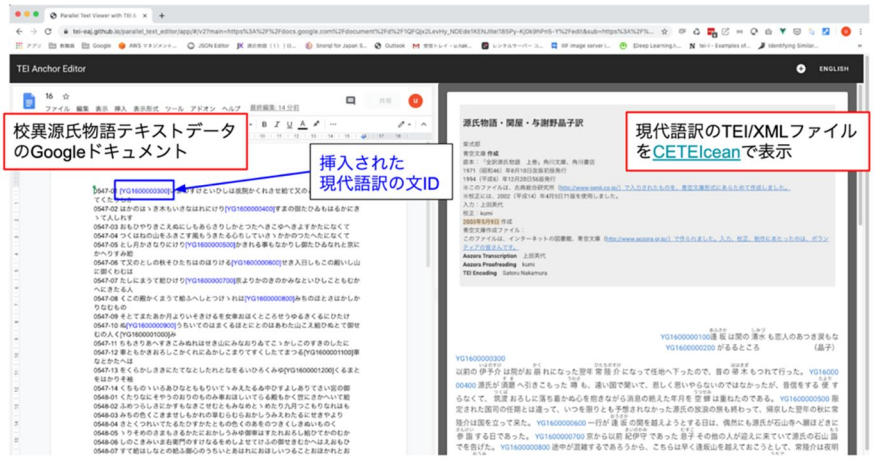

図 4 <anchor/>タグを用いたテキストの関連づけ

4．研究成果

3 で作成したデータを用いて構築したシステム「デジタル源氏物語」について述べる．デジタル源氏物語のシステム概要図を図 5 に示す．校異源氏物語のテキストデータを公開する「校異源氏物語テキスト DB」と，各種データを関連づけて公開する「デジタル源氏物語」の 2 種類のアプリケーションから構成される．以下，これらについて述べる．

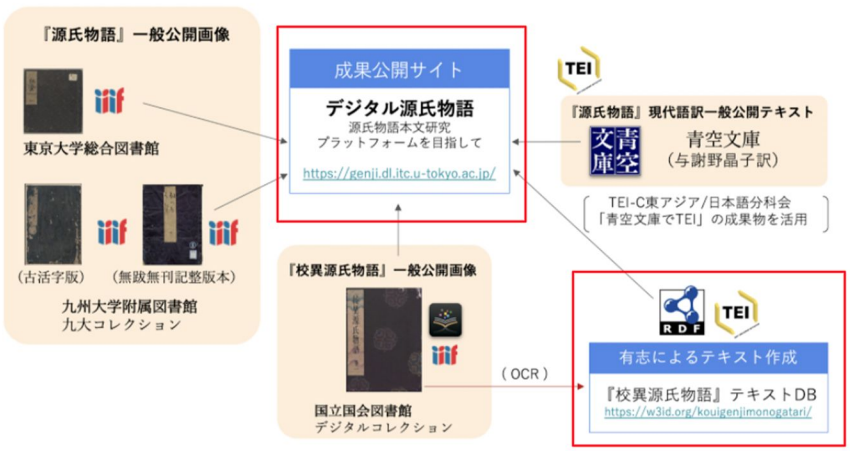

図 5 デジタル源氏物語のシステム概要図

4.1 校異源氏物語テキスト DB[4]

本ウェブアプリケーションは，校異源氏物語テキストの TEI/XML ファイルと，行情報の RDF データを提供し，以下の 3 つの機能を提供する．

1 点目は，図 6 に示すように，TEI テキストと IIIF 画像を並列に表示する機能である．2 点目は，TEI/XML ファイルをそのまま提供する機能であり，本サイトをホスティングしている GitHub リポジトリに遷移する．3 点目は，行情報の RDF データを提供する機能である．本データは CC0 のライセンスで提供しているため，様々用途に利用することができる．

図 6 TEI テキストと IIIF 画像の並列表示

4.2 デジタル源氏物語

本ウェブアプリケーションは，3 で作成した 3 種類のデータを関連づけて提供する．その代表的な機能が，図 1 に示した機能である．その他，校異源氏物語の頁数毎に画像を比較する機能や，IIIF 対応の源氏物語のリストを提供する．

4.3 まとめ

本研究では，オンライン上で複数の利用者が史料を翻刻可能なシステムを開発した．特に，画像共有のための国際標準である IIIF や，人文学資料向けの構造化ルールを定める TEI 等の国際規格に適合させることで，幅広い用途・国際的に活用可能なシステムを構築した．本システムを用いて，画像と多様なテキストデータを関連付け，「源氏物語」の本文研究支援を目指すウェブサイト「デジタル源氏物語」を公開した．

オンラインでの協力的な史料翻刻システムの開発により，人文情報学分野の発展に寄与する．また多様なテキストデータと画像を統合することで，「デジタル源氏物語」ウェブサイトの公開など，学術研究基盤の強化に貢献した．さらにくずし字 OCR の利用，およびテキストデータが類似する写本・版本の画像を自動的に推奨する機能の組み合わせにより，歴史資料の新しい活用方法を提案した．これらの取り組みは，歴史資料へのアクセスを容易とし，国際的な学術研究に寄与することが期待される．

参考文献

[1]．“ デジタル源氏物語 ”．https://genji.dl.itc.u-tokyo.ac.jp/，(参照 2023-05-30)．
[2]．“ 青空文庫 源氏物語 “ ．https://www.aozora.gr.jp/cards/000052/card362.html，(参照 2020-07-26)．
[3]．“ KuroNet くずし字認識サービス ”，http://codh.rois.ac.jp/kuronet/，(参照 2023-05-30)．
[4]．“ 校異源氏物語テキスト DB ”，https://kouigenjimonogatari.github.io/，(参照 2023-05-30)．

５．主な発表論文等

〔雑誌論文〕　計14件（うち査読付論文　10件／うち国際共著　1件／うちオープンアクセス　2件）

１．著者名 中村 覚，田村 隆，永崎 研宣	４．巻 2022-CH-128(13)
２．論文標題 デジタル源氏物語（AI画像検索版）：くずし字OCRと編集距離を用いた写本・版本の比較支援システムの開発	５．発行年 2022年
３．雑誌名 研究報告人文科学とコンピュータ（CH）	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 劉 冠偉，中村 覚，山田 太造	４．巻 2022-CH-128(2)
２．論文標題 部品と画数で漢字を検索するためのUnicode入力支援ツール	５．発行年 2022年
３．雑誌名 研究報告人文科学とコンピュータ（CH）	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 鳥居 克哉，中村 覚，山田 太造，稗方 和夫	４．巻 2022-CH-128(8)
２．論文標題 日本中世古記録を対象としたトピック抽出自動化システムの構築	５．発行年 2022年
３．雑誌名 研究報告人文科学とコンピュータ（CH）	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 小風 尚樹，中村 覚，永崎 研宣，渡辺 美紗子，戸村 美月，小風 綾乃，清武 雄二，後藤 真，小倉 慈司	４．巻 2021
２．論文標題 相互運用性を高めた日本歴史資料データ実装:『延喜式』TEI と IIIF を事例として	５．発行年 2021年
３．雑誌名 じんもんこん2021論文集	６．最初と最後の頁 294-301
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 橋本雄太，金甫榮，中村覚，小風尚樹，井上さやか，茂原暢，永崎研宣	4．巻 2021
2．論文標題 写真資料のクラウドアノテーションシステムの開発: 『渋沢栄一伝記資料」別巻第 10 を事例に	5．発行年 2021年
3．雑誌名 じんもんこん2021論文集	6．最初と最後の頁 132-137
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 中村覚，須田牧子，黒嶋敏，井上聡，山田太造	4．巻 2021
2．論文標題 データ駆動型歴史情報研究基盤の構築に向けた知識ベースの構築とその活用: 絵図史料を対象として	5．発行年 2021年
3．雑誌名 じんもんこん2021論文集	6．最初と最後の頁 88-95
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 山田太造，中村覚，渋谷綾子，大向一輝，井上聡	4．巻 2021
2．論文標題 日本史史料を対象とした研究データ基盤整備における課題	5．発行年 2021年
3．雑誌名 じんもんこん2021論文集	6．最初と最後の頁 80-87
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Satoru Nakamura, Taizo Yamada	4．巻 11
2．論文標題 Development of data-driven historical information research infrastructure at the Historiographical Institute in the University of Tokyo	5．発行年 2021年
3．雑誌名 The 11th International Conference of Japanese Association for Digital Humanities	6．最初と最後の頁 148-151
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Satoru Nakamura, Ayano Kokaze, Yoshiho Iida, Naoki Kokaze, Tatsuo Hemmi	4．巻 11
2．論文標題 Development of a support system for extracting mentioned bibliographical data from the Encyclop_die entries	5．発行年 2021年
3．雑誌名 The 11th International Conference of Japanese Association for Digital Humanities	6．最初と最後の頁 130-133
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Boyoung Kim, Satoru Nakamura, Yuta Hashimoto, Naoki Kokaze, Sayaka Inoue, Toru Shigehara, Kiyonori Nagasaki	4．巻 11
2．論文標題 Reconstruction and Utilization of Text Data Using TEI: Case study of the Shibusawa Eiichi Denki Shiryo	5．発行年 2021年
3．雑誌名 The 11th International Conference of Japanese Association for Digital Humanities	6．最初と最後の頁 126-129
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Satoru Nakamura	4．巻 2021
2．論文標題 The University of Tokyo Digital Archives Development Project: Developing an Approach for Utilizing Academic Assets across Different Organizations	5．発行年 2020年
3．雑誌名 The National Museum of Japanese History. Japanese and Asian Historical Research In the Digital Age	6．最初と最後の頁 17～36
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 金 甫榮，中村 覚，小風 尚樹，橋本 雄太，井上 さやか，茂原 暢，永崎 研宣	4．巻 2020
2．論文標題 TEIを用いた『渋沢栄一伝記資料』テキストデータの再構築	5．発行年 2020年
3．雑誌名 じんもんこん2020論文集	6．最初と最後の頁 47～52
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 中村 覚、高嶋 朋子	4．巻 5
2．論文標題 持続性と利活用性を考慮したデジタルアーカイブ構築手法の提案	5．発行年 2021年
3．雑誌名 デジタルアーカイブ学会誌	6．最初と最後の頁 56〜60
掲載論文のDOI（デジタルオブジェクト識別子） 10.24506/jsda.5.1_56	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 中村覚，田村隆，永崎研宣	4．巻 2020-CH-124
2．論文標題 源氏物語本文研究支援システム「デジタル源氏物語」の開発におけるIIIF・TEIの活用	5．発行年 2020年
3．雑誌名 研究報告人文科学とコンピュータ（CH）	6．最初と最後の頁 1〜7
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 中村覚
2．発表標題 東京大学デジタルアーカイブズ構築事業の取り組みとその利活用について
3．学会等名 2020年度KU-ORCAS国際シンポジウム：デジタルヒューマニティーズ推進のための環境構築とその課題
4．発表年 2021年

1．発表者名 中村覚
2．発表標題 IIIF Curation Platformを用いたデジタルアーカイブの活用
3．学会等名 第14回CODHセミナー：IIIF Curation Platform利活用レシピ100連発
4．発表年 2021年

１．発表者名
中村覚

２．発表標題
源氏物語本文研究支援システム「デジタル源氏物語」の開発におけるIIIFとTEIの活用

３．学会等名
U-PARL [協働型アジア研究オンラインセミナー]IIIFに準拠した画像公開の方法とTEIとの連携

４．発表年
2021年

１．発表者名
田村隆，中村覚，中村美里，永崎研宣

２．発表標題
「デジタル源氏物語」の構築と展開

３．学会等名
国文学研究資料館 第6回日本語の歴史的典籍国際研究集会

４．発表年
2020年

１．発表者名
中村覚

２．発表標題
デジタルアーカイブ活用のために 最新技術の紹介

３．学会等名
第3回東京大学学術資産アーカイブ化推進室主催セミナー

４．発表年
2019年

１．発表者名
田村隆

２．発表標題
東大本『源氏物語』と新たな本文研究プラットフォーム

３．学会等名
第3回東京大学学術資産アーカイブ化推進室主催セミナー

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

倭寇図巻デジタルアーカイブ https://www.hi.u-tokyo.ac.jp/collection/degitalgallary/wakozukan/ デジタル源氏物語（AI画像検索版） https://genji-ai.web.app/ 渋沢栄一ダイアリー https://shibusawa-dlab.github.io/app1/ デジタル延喜式 https://khirin-t.rekihaku.ac.jp/engishiki/ デジタル源氏物語 https://genji.dl.itc.u-tokyo.ac.jp/ 校異源氏物語テキストDB https://kouigenjimonogatari.github.io/ デジタル源氏物語（AI画像検索版） https://genji-ai.web.app/ デジタル源氏物語 https://genji.dl.itc.u-tokyo.ac.jp/app/#/ 校異源氏物語テキストDB https://kouigenjimonogatari.github.io/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関