Vol.2020-CH-124 No.2
2020/9/5

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

中村覚[1]　田村隆[1]　永崎研宣[2]

概要：「デジタル源氏物語」は，『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．本発表では，特に本システム開発におけるデータ構築とアプリケーション構築について述べる．具体的には，TEI を用いたテキストデータの作成や現代語訳との関連付け，IIIF を用いたくずし字 OCR の活用やテキストデータとの関連づけなどの手法について述べ，その有用性を検証する．

キーワード：源氏物語，IIIF，TEI，LOD，くずし字

1. はじめに

1.1 背景

「デジタル源氏物語[1]」は，『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．東京大学総合図書館所蔵『源氏物語[2]』の公開（2019 年 6 月）を契機に，有志により「『源氏物語』研究にとって有意義なデジタル機能は何か」という検討から開始した．

人文情報学的な手法を用いた源氏物語に関する研究は数多く存在し，例えば宮脇文経氏による「源氏物語の世界再編集版[3]」がある．本サイトは高千穂大学の渋谷名誉教授が公開しているホームページ「源氏物語の世界[4]」を再編集したもので，注釈の可読性の向上，本文と現代語訳の対照表示機能の提供などの特徴を持つ．本文や注釈，現代語訳等を XML ファイルとして記述し，これらを関連づけることで対照表示を可能としている．これらの成果を活用しつつ，本研究は IIIF や TEI 等の国際標準規格を採用し，また画像データとテキストデータの関連づけを行う点に差異がある．

IIIF や TEI 等の国際標準規格を採用し，また画像データとテキストデータの関連づけを行うことにより，研究支援システムを開発している例として，永崎ら[5][6]や高橋ら[7]の研究がある．永崎らは，仏教研究のための協働研究プラットフォームとしての大蔵経データベースを 2008 年より構築・公開しており，2018 年版では TEI と IIIF を採り入れ、TEI テキストや世界各地の IIIF 対応画像と本文との対応づけ作業を Web 上で実施できる協働作業環境を提供している．また高橋らは文字研究・言語研究のためのプラットフォームシステムを構築しており，IIIF アノテーションを用いた画像データの効率的な利用環境と，TEI・LOD を用いた他者とのデータ共有環境を実現している．本研究では，これらの研究における IIIF や TEI の活用方法を参考としつつ，源氏物語の本文研究の支援に適したシステムの開発を目指している．

1.2 デジタル源氏物語の提供機能例

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[8]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 Mirador 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て表示される．

1 東京大学
The University of Tokyo.
2 人文情報学研究所
International Institute for Digital Humanities.

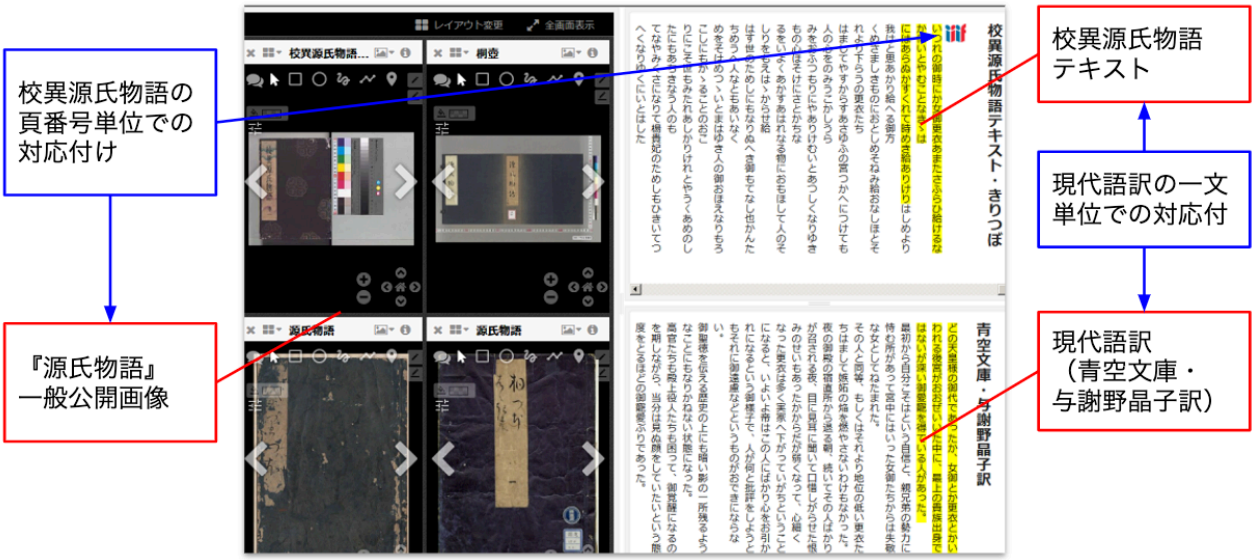

図 1　デジタル源氏物語の提供機能例

Figure 1　Example of a function provided by “Digital The Tale of Genji”.

この機能の実現にあたっては，以下のデータが必要となる．

まずは，校異源氏物語のテキストデータである．IIIF に準拠した画像が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ため，この画像中の文字列のテキスト化が求められる．

次に，源氏物語の一般公開画像である．今日では，東京大学や九州大学，国文学研究資料館をはじめとして，様々な機関から源氏物語の画像が IIIF に準拠して公開されている．ただし，これらの画像には校異源氏物語の頁番号が付与されていないことが一般的であり，各画像と頁番号の対応づけが必要となる．なお，九州大学では一部のコレクションについて「対応頁検索」を提供しており，源氏物語大成の頁数（校異源氏物語と同頁）が古活字版と無跋無刊記整版本にすでに付与されている[9]（図 2）．

最後に，青空文庫で公開されている与謝野晶子による現代語訳である．特に，上述した対応箇所のハイライト表示にあたっては，校異源氏物語のテキストデータと現代語訳を文などの単位で対応づける必要がある．

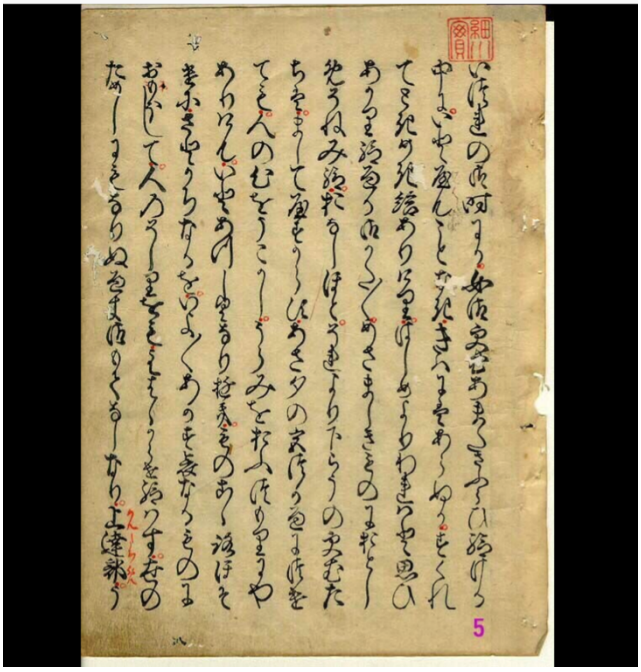

図 2　源氏物語大成の頁数（校異源氏物語と同頁）の付与例

Figure 2　Example of the number of pages of the “源氏物語大成”.

1.3 目的

上述した必要要件に基づき，本研究ではデジタル源氏物語のシステム構築におけるデータ構築と，それらを提供するためのアプリケーション構築について述べる．具体的には，TEI を用いたテキストデータの作成や現代語訳との関連付け，IIIF を用いたくずし字 OCR の活用やテキストデータとの関連づけなどの手法について述べ，その有用性を検証する．

なお，デジタル源氏物語の人文学研究における意義や，提供機能の有用性については，田村氏の論考[10]等を参照されたい．

2. データ構築

先述した以下のデータ構築作業について，各々説明する．

- 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

2.1 校異源氏物語のテキストデータ作成

本作業内容としては，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[11]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 3 にその画面例を示す．

図 3　Scripto を用いた翻刻画面の例

Figure 3　Example of a transcription interface provided by “Scripto”.

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ことができる．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．

具体的には，以下の手順でデータ構築を行った．

1. 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の IIIF 画像を使用して，巻毎の IIIF マニフェストを生成
2. Omeka S の CSV Import モジュールを使って一括登録
   (ア) IIIF 画像 URL を指定
3. MediaWiki API を使用して，OCR テキストを一括登録
4. Scripto を使用して，テキストデータを修正
5. Omeka S と MediaWiki の API を使用して，テキストデータの取得，TEI/XML 形式に変換して保存
   (ア) Best Practices for TEI in Libraries における Level 2 程度のマークアップ[12]
   (イ) IIIF マニフェストの対応付け[13]
   (ウ) 各行に URI を付与し，ジャパンサーチの利活用スキーマを参考にした RDF データを作成

2.2 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，以下に示す手順で実施した．

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[14]」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．KuroNet くずし字認識サービスでは，2020 年 3 月から「自動読み順推定アルゴリズム」を提供している．これを利用することで，図 4 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．

次に，くずし OCR によって作成したテキストデータと，2.1 で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．

図 4　校異源氏物語の頁数の自動付与
Figure 4　Automatically assign the number of pages of the “校異源氏物語”.

また，校異源氏物語の頁数付与に加えて，新編日本古典文学全集の頁数も合わせて付与した．この付与にあたっては，古典研究所で公開されているテキストデータ[15]を利用することで，校異源氏物語の頁数付与と同じプロセスで進めることができた．

これらの結果をデータ管理用として IIIF Curation，表示用としてアノテーションリスト付きの IIIF マニフェストの形式で保存した．この理由として，本プロジェクトでは複数の機関が公開する IIIF マニフェストに関する情報を管理する必要があり，この時，IIIF Curation ではひとつのファイル内で複数の IIIF マニフェストを扱うことでできる点に利点がある．複数の IIIF マニフェストを管理する方法としては IIIF コレクションも挙げられるが，特定の canvas 内の特定の箇所を指定した上での管理には適していない．また，データ表示用にはアノテーションリストを作成することで，図 5 に示すようなアノテーション表示を行った．これにより，校異源氏物語，新編日本古典文学全集の頁数から，国立国会図書館デジタルコレクション，ジャパンナレッジで公開されている各資料のデジタル画像へ遷移することを可能とした．加えて，Mirador ビューアを用いた，同一画面上での複数画像の比較も可能となる．

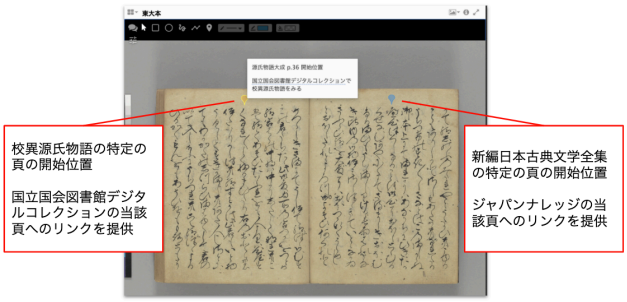

図 5　校異源氏物語，新編日本古典文学全集の頁数に関するアノテーション例
Figure 5　Example of annotations.

2.3 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．この際，SAT 大蔵経データベース 2018 年版[16]で提供している現代日本語訳と漢訳の対応づけ表示システムに倣い，文単位で ID を付与した．

次に，2.1 で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 6 に示す，本作業を支援するウェブアプリケーション[17]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 CETEIcean で表示する．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．

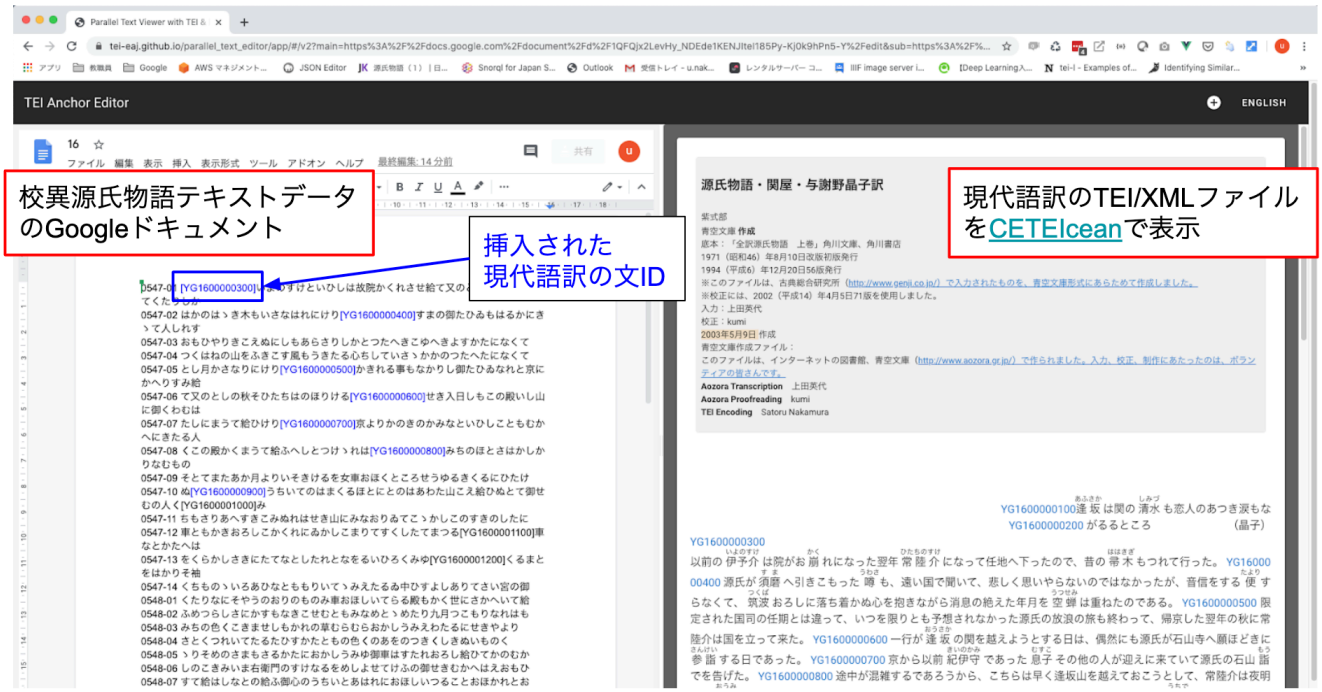

図 6　<anchor/>タグを用いたテキストの関連づけ

Figure 6　Text linking with <anchor/> element.

2.4 まとめ

これらの作業を通じて，以下のデータを構築した．

- 校異源氏物語のテキストデータ作成：TEI 準拠の XML ファイル
- 公開画像への校異源氏物語の頁数付与：IIIF Curation API 準拠の JSON ファイル
- 校異源氏物語と現代語訳の対応付け：TEI 準拠の XML ファイル

3. アプリケーション構築

ここでは，2 で作成したデータを用いて構築したアプリケーションについて述べる．デジタル源氏物語のシステム概要図を図 7 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．

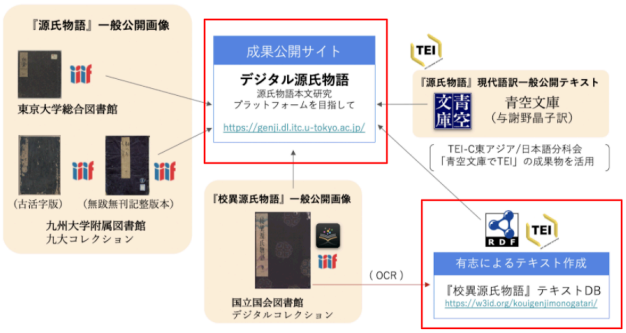

図 7　デジタル源氏物語のシステム概要図

Figure 7　System overview.

3.1 校異源氏物語テキスト DB[18]

本サイトは，校異源氏物語テキストの TEI/XML ファイルと，行情報の RDF データを提供する．以下の 3 つの機能を提供する．

1 点目は，TEI-C 東アジア／日本語分科会が作成している TEI Multi Viewer[19]を使用して，図 8 に示すように，TEI テキストと IIIF 画像を並列に表示する．

図 8　TEI テキストと IIIF 画像の並列表示
Figure 8　Parallel view with TEI text and IIIF image.

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り，神崎正英氏が作成している Linked Data Browser[20]に遷移する．

さらに本 GitHub リポジトリでは，教科書 LOD[21]における取り組みを参考として，w3id.org を利用した行 URI の固定化を実施している．

本データは CC0 のライセンスで提供しているため，様々用途にご利用いただければ幸いである．

3.2 デジタル源氏物語

本サイトは，2 で作成した 3 種類のデータを関連づけて提供するサイトである．その代表的な機能が，図 1 に示した機能である．この各種データを組み合わせた表示にあたっては，TEI-C 東アジア／日本語分科会が作成している TEI Parallel Text Viewer[22]を使用している．なお，これらのデータ作成作業は進行中のため，作業の進捗状況をサイト上で提示している．

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している．

今後もデータ追加や機能追加を検討している．

4. 考察

ここでは，データ構築・アプリケーション構築手法における考察を行う．

4.1 マイクロサービスアーキテクチャの採用

本データ構築・アプリケーション構築にあたっては，各種サービスを組み合わせて利用するマイクロサービスアーキテクチャに準じた手法を採用した．

データ構築にあたっては，IIIF，Omeka S（REST API, Scripto, MediaWiki），IIIF Curation Platform，くずし字 OCR，Google ドキュメントなどのサービスを適宜利用した．これらのサービスが各々提供する機能を活用し，それらの API を介して関連づけることにより，データ構築作業を効率的に行うことができた．

アプリケーション構築においては，作成した IIIF, TEI, RDF データについて，TEI-C 東アジア／日本語分科会が作成している各種 TEI 対応ビューアや，神崎正英氏が作成している Linked Data Browser を利用した．デジタル源氏物語のようなデータセットの提供に主眼を置いたプロジェクトにおいては，標準規格に準拠した形式でデータを作成し，既存の各種アプリケーションを使用する手法を採用することで，既存の成果を活用したデータ提供を簡易に実現しつつ，本プロジェクトが独自に提供するアプリケーション部分の構築に注力するができた．

4.2 くずし字 OCR と編集距離を用いた校異源氏物語・新編日本古典文学全集の頁番号の自動付与

本研究では，2.2 で述べた通り，くずし字 OCR と編集距離を用いた校異源氏物語・新編日本古典文学全集の頁番号の自動付与を行った．これにより，源氏物語研究の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を可能とすることで，本作業に要するコストを大幅に削減することができた．

この点について，源氏物語研究の専門家からくずし字 OCR の有用性が高く評価されるとともに，「OCR 結果の不完全さを補う手法として編集距離は有益である」というコメントをいただいた．このような取り組みが，くずし字 OCR の利活用の発展に寄与することができれば幸いである．

5. 結論

本研究では，『デジタル源氏物語』システムにおけるデータ構築とアプリケーション構築について述べた．具体的には，TEI を用いたテキストデータの作成や現代語訳との関連付け，IIIF を用いたくずし字 OCR の活用やテキストデータとの関連づけなどの手法について述べ，その有用性を検証した．

今後は本システムの研究利用を通じて，人文学研究における意義や有用性について検討していきたい．また本手法の汎用性を生かし，湖月抄などを追加登録する予定であり，引き続き対象画像の拡大を進めていく．さらに本手法をモジュール化することで，第三者がデジタル源氏物語にデータを追加可能な仕組みや，源氏物語以外にも応用しやすい仕組みを構築していきたい．

謝辞　本プロジェクトにご協力いただいた関係者のみなさま，特に東京大学総合図書館および情報システム部の職員のみなさまに感謝申し上げます．またくずし字 OCR の

利用にあたっては, CODH 北本朝展先生, カラーヌワット・タリン先生にご協力いただきました. 深く感謝いたします.

本研究は JSPS 科研費 19K20626 の助成による成果の一部です.

参考文献

[1] “デジタル源氏物語”. https://genji.dl.itc.u-tokyo.ac.jp/, (参照 2020-07-26).
[2] “東京大学総合図書館所蔵『源氏物語』“. https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/genji/page/home, (参照 2020-07-26).
[3] “源氏物語の世界 再編集版”. http://www.genji-monogatari.net/, (参照 2020-07-26).
[4] “源氏物語の世界”. http://www.sainet.or.jp/~eshibuya/index.html, (参照 2020-07-26).
[5] Kiyonori Nagasaki, A. Charles Muller, Toru Tomabechi, Masahiro Shimoda. A Collaborative System for Digital Research Environment via IIIF, Digital Humanities 2019, https://dev.clariah.nl/files/dh2019/boa/0378.html, (参照 2020-08-11).
[6] Kiyonori Nagasaki, Ikki Ohmukai, Toru Tomabechi, Masahiro Shimoda. An Improvement of Collaborative Digital Scholarly Edition with IIIF, Digital Humanities 2020, https://dh2020.adho.org/wp-content/uploads/2020/07/515_AnImprovementofCollaborativeDigitalScholarlyEditionwithIIIF.html, (参照 2020-08-11).
[7] 高橋洋成, 永井正勝, 和氣愛仁. 画像, TEI, LOD を用いた文字研究・言語研究のためのプラットフォームの構築, 研究報告人文科学とコンピュータ（CH）, Vol.2015-CH-105, No.5, pp.1-5, 2015.
[8] “青空文庫 源氏物語“. https://www.aozora.gr.jp/cards/000052/card362.html, (参照 2020-07-26).
[9] “デジタル化画像 - 貴重資料（九大コレクション）”, https://guides.lib.kyushu-u.ac.jp/rare/images, (参照 2020-08-11).
[10] 田村隆. 東大本『源氏物語』と新たな本文研究プラットフォーム. 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, 2019.11, http://hdl.handle.net/2261/00078929
[11] “Scripto”. https://omeka.org/s/modules/Scripto/, (参照 2020-07-26).
[12] “Best Practices for TEI in Libraries”, http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html, (参照 2020-07-27).
[13] “IIIF 画像とのリンク · TEI-EAJ/jp_guidelines Wiki”, https://github.com/TEI-EAJ/jp_guidelines/wiki/IIIF%E7%94%BB%E5%83%8F%E3%81%A8%E3%81%AE%E3%83%AA%E3%83%B3%E3%82%AF, (参照 2020-07-27).
[14] “KuroNet くずし字認識サービス”, http://codh.rois.ac.jp/kuronet/, (参照 2020-07-27).
[15] “「源氏物語」新編日本古典文学全集”, http://www.genji.co.jp/zenshu-genji-srch.php, (参照 2020-07-29).
[16] “SAT 大蔵経データベース 2018 年版”, https://21dzk.l.u-tokyo.ac.jp/SAT2018/master30.php,（参照 2020-08-11）.
[17] “TEI Anchor Editor”. https://tei-eaj.github.io/parallel_text_editor/app/#/v2, (参照 2020-07-27).
[18] “校異源氏物語テキスト DB”, https://kouigenjimonogatari.github.io/, (参照 2020-07-27).
[19] “TEI Multi Viewer”, https://tei-eaj.github.io/tei_viewer/app/, (参照 2020-07-27).
[20] “Linked Data Browser”, https://www.kanzaki.com/works/2014/pub/ld-browser, (参照 2020-07-27).
[21] “教科書 LOD”, https://jp-textbook.github.io/, (参照 2020-07-27).
[22] “TEI Parallel Text Viewer”, https://tei-eaj.github.io/parallel_text_viewer/, (参照 2020-07-27).