

「デジタル源氏物語」を 支える技術: IIIFとTEI

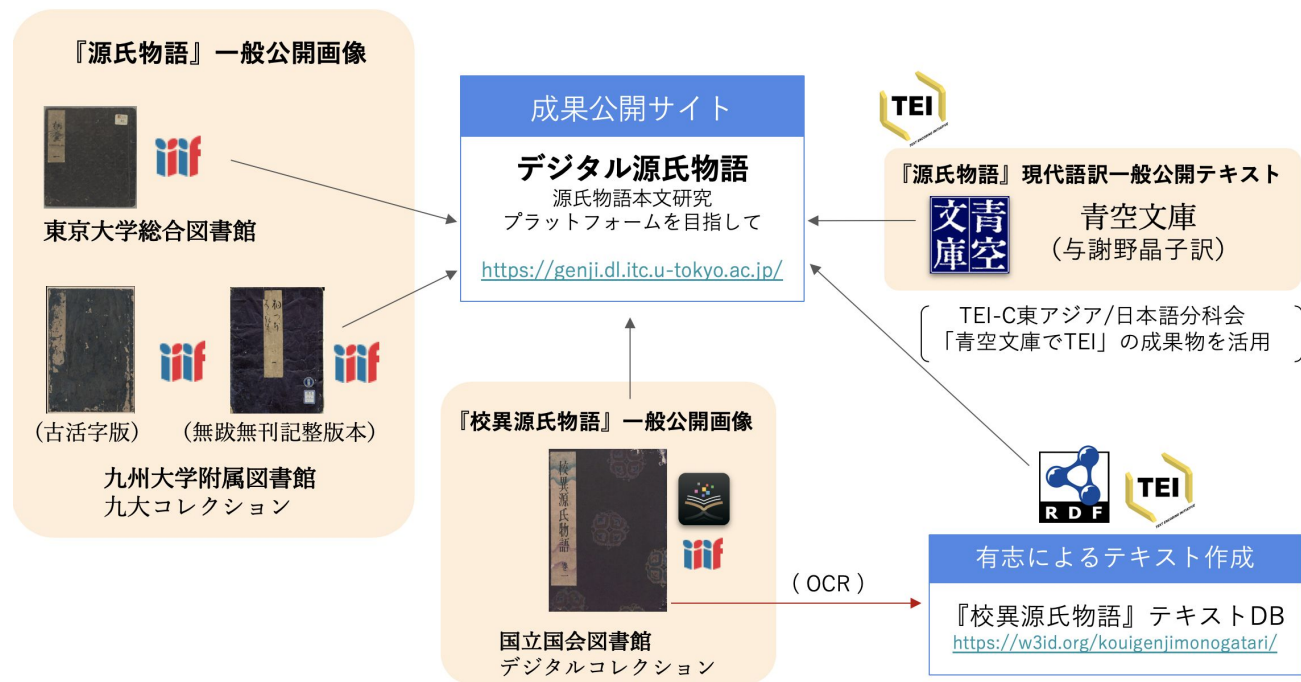
中村 覚
東京大学

もくじ

- 背景・目的
- データ構築
- 結論

デジタル源氏物語

『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指したシステム



校異源氏物語：
『源氏物語』主要
本文の校異を示
した研究書

『源氏物語』
一般公開画像

現代語訳の一文単位での対応付

現代語訳
(青空文庫・
与謝野晶子訳)



要素技術 (TEI・IIF) の話 =>

歴史情報の利活用にむけて

提供

- デジタルアーカイブ
 - 持続可能性
 - OSS (Open Source Software)
- DOI (Digital Object Identifier)
- オープンライセンス
 - CC License
 - PDM (Public Domain Mark)



提供者と利用者をつなぐ規格・技術

メタデータ



LOD (Linked Open Data)
RDF

画像



IIF
(International Image
Interoperability Framework)

テキスト



TEI
(Text Encoding Initiative)

⋮

利用

- 研究・教育
 - 機械学習
 - 人文情報学
 - オープンサイエンス
 - **くずし字OCR**
- 普及・啓蒙活動
 - 展示会
 - 電子展示
- 学術情報の流通
 - 機関の横断
 - 分野の横断

TEIとは

- TEI (Text Encoding Initiative)協会が作っているTEIガイドラインを指すことが多い。
- 1987年に始まる人文学研究者・情報系研究者と図書館司書による電子テキストの効果的効率的な共有のためのコミュニティ活動
 - 技術委員会を作ってメンバーをコミュニティで選任し、「ガイドライン」をアップデートし続けている。
 - 現在はテキスト以外も対象にしている。

TEIガイドラインとは

- 人文学向け資料を共有しやすいように構造化(現時点ではXML化)するためのルール
- 言語学・文献学・文学研究向けのルールセット
- 研究用ツール作成のためのルールも含む
 - 辞書、図形、外字、詳細な書誌情報、
- 「どの情報は誰に責任があるか」「どの程度あてになるか」の記述
- 一度書けば色々なアプリで活用できるように
- 研究に活用できる要素をうまく抽出・記述
- 専門家が付与した詳細情報が永続的に使えるように
 - ソフトが変わって使えなくなった、ということがないように

TEI/XMLの例

← 書誌情報の記述例

(永崎研宣氏作成)

↓ 本文のマークアップ例

```
<titleStmt>
  <title>走れメロス</title>
  <author>太宰治</author>
  <respStmt>
    <resp>Aozora Transcription</resp>
    <name>金川一之</name>
  </respStmt>
  <respStmt>
    <resp>Aozora Proofreading</resp>
    <name>高橋美奈子</name>
  </respStmt>
  <resp>
    <resp when="2011-01-17">作成</resp>
    <orgName ref="http://www.aozora.gr.jp/">青空文庫</orgName>
  </resp>
  <note>
    <lb/>底本：「太宰治全集3」ちくま文庫、筑摩書房
    <lb/>1988（昭和63）年10月25日初版発行
    <lb/>1998（平成10）年6月15日第2刷
    <lb/>底本の親本：「筑摩全集類聚版太宰治全集」筑摩書房<lb/>
    <lb/>1975（昭和50）年6月～1976（昭和51）年6月
    <lb/>入力：金川一之
    <lb/>校正：高橋美奈子
  </note>
</title>
```

```
<body>
  <p>
    <persName corresp="#メロス">メロス</persName>は激怒した。必ず、かの
    <persName corresp="#ディオニス">邪智暴虐（じゃちぼうぎやく）
      の王</persName>を除かなければならぬと決意した。
    <persName corresp="#メロス">メロス</persName>には政治がわからぬ。
    <persName corresp="#メロス">メロス</persName>は、村の牧人である。
    笛を吹き、羊と遊んで暮して来た。けれども邪悪に対しては、人一倍に敏感であった。
  </p>
</body>
```

視覚化アプリの例

TEI Level 4 Viewer

人物情報

ディオニス
王

外部サイト

「ディオニス」を調べる

 Wikipedia

 Twitter

 Amazon

走れメロス

太宰治

メロス 牧人

ディオニス 王

セリヌンティウス 石工

メロスは激怒した。必ず、かの邪智暴虐の王を除かなければならぬと決意した。メロスには政治がわからぬ。メロスは、村の牧人である。笛を吹き、羊と遊んで暮して来た。けれども邪悪に対しては、人一倍に敏感であった。

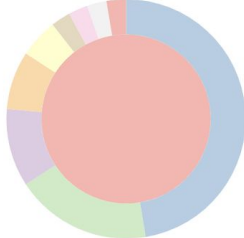
きょう末明メロスは村を出発し、野を越え山越え、十里はなれた此のシラクスの市にやって来た。メロスには父も、母も無い。女房も無い。十六の、内気な妹と二人暮らしだ。この妹は、村の或る律気な一牧人を、近々、花婿として迎える事になっていた。結婚式も間近かなのである。メロスは、それゆえ、花嫁の衣裳やら祝宴の御馳走やらを買いに、はるばる市にやって来たのだ。先ず、その品々を買い集め、それから都の大路をぶらぶら歩いた。

メロスは竹馬の友があった。セリヌンティウスである。今は此のシラクスの市で、石工をしている。その友を、これから訪ねてみるつもりなのだ。久しく逢わなかったのだから、訪ねて行くのが楽しみである。

歩いているうちにメロスは、まちの様子を怪しく思った。ひっそりしている。もう既に日も落ちて、まちの暗いのは当りまえだが、けれども、なんだか、夜のせいばかりでは無く、市全体が、やけに寂しい。のんきなメロスも、だんだん不安になって来た。路で逢った若い衆をつかまえて、何かあったのか、二年まえに此の市に来たときは、夜でも皆が歌をうたって、まちは賑やかであった筈だが、と質問した。若い衆は、首を振って答えなかった。しばらく歩いて老舗に逢い、こんどはもっと、語勢を強くして質問した。老舗は答えなかった。メロスは両手で老舗のからだをゆすぶって質問を重ねた。老舗は、あたりをはばかりの低声で、わずか答えた。「王様は、人を殺します。」「なぜ殺すのだ。」「悪心を抱いている、

「ディオニス」の呼称 (38件)

グラフを隠す



発話内容

「ディオニス」の発話 (10件)

ワードクラウドを表示

- 「この短刀で何をするつもりであったか。言え！」
- 「おまえがか？」
- 「仕方の無いやつじゃ。おまえには、わ

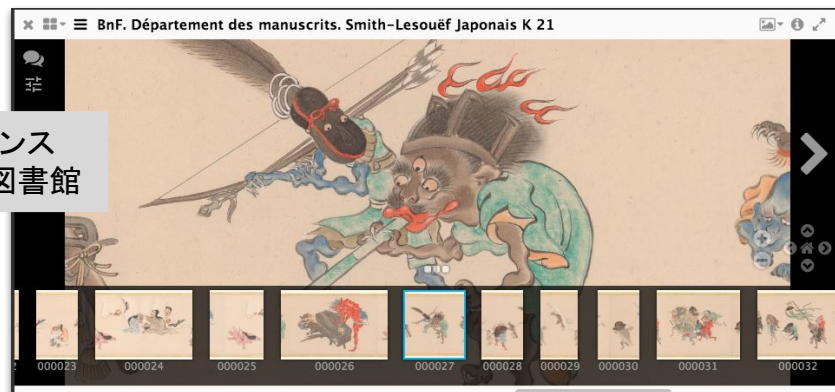
IIIF (トリプルアイエフ) とは？



- International Image Interoperability Framework
- 画像などのWebコンテンツを共有するための国際的な枠組み
- 2010年頃から、スタンフォード大図書館、英国図書館、フランス国立図書館、オックスフォード大ボドリアン図書館などの主導により開発が進められる。
- 現在、国内外600機関以上がIIIFを採用し、画像を公開している。
- おもな導入機関(国内)
 - 東京大学、京都大学、島根大学、九州大学、慶應義塾大学、国立国会図書館、国文学研究資料館、国立歴史民族博物館、etc...

複数の機関から公開されている画像を1つのビューアで表示している例:『百鬼夜行図』

フランス
国立図書館



東京大学
総合図書館



国立国会
図書館



国文学研究
資料館

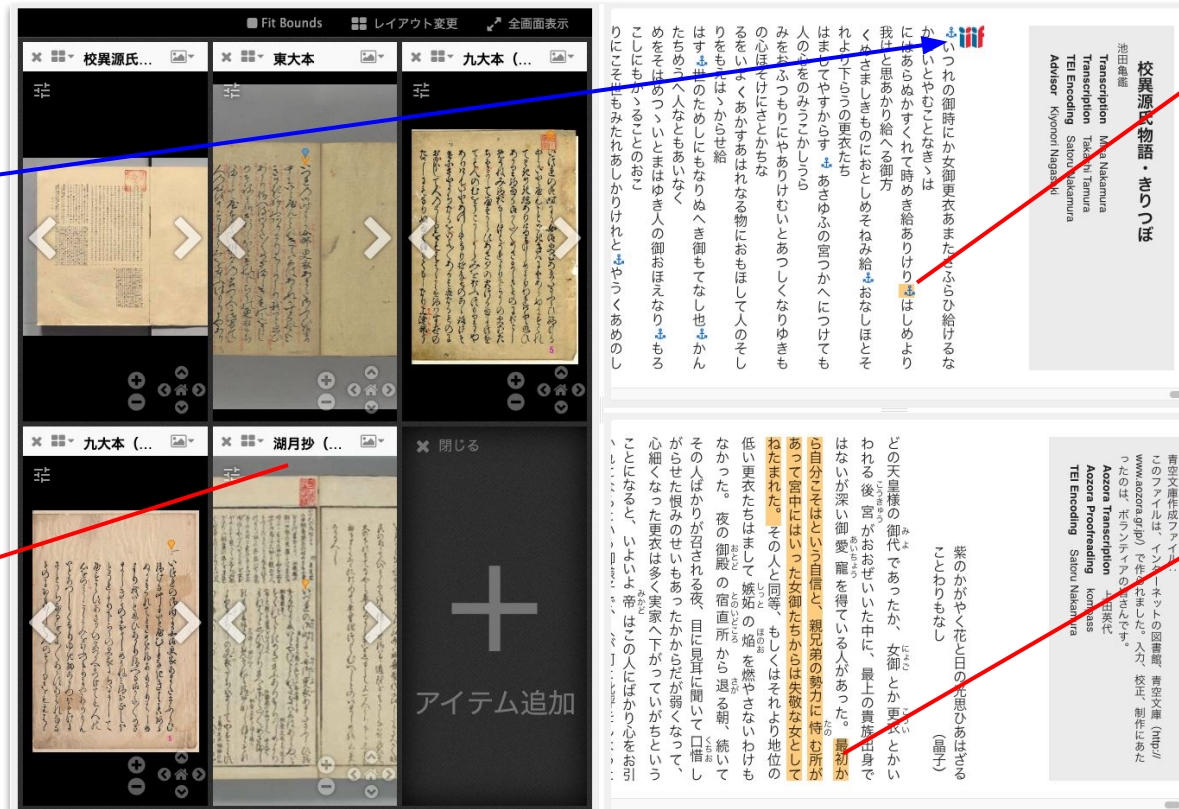


デジタル源氏物語の話に戻ります =>

デジタル源氏物語の提供機能例

校異源氏物語の
頁番号単位での
対応付け

『源氏物語』
一般公開画像



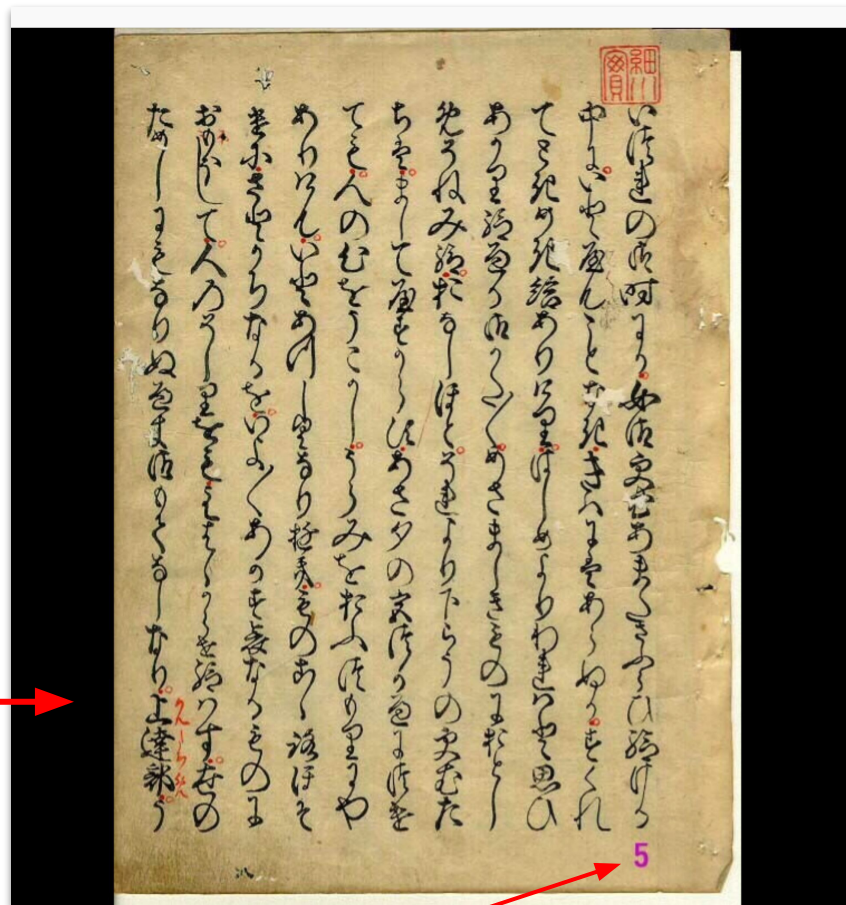
校異源氏物語
テキスト

現代語訳の一文
単位での対応付

現代語訳
(青空文庫・
与謝野晶子訳)

機能提供に必要なデータ

- 校異源氏物語のテキストデータ
 - IIF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
- 『源氏物語』一般公開画像
 - 東京大学、九州大学、国文学研究資料館をはじめ、様々な機関からIIF画像が公開されている
 - 各画像と校異源氏物語の頁番号を対応づける必要あり
- 現代語訳(青空文庫・与謝野晶子訳)
 - 校異源氏物語のテキストデータと対応づける必要あり



本発表の目的

- 『デジタル源氏物語』システムにおけるデータ構築について述べる。
 - **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

データ構築

データ構築

① 校異源氏物語のテキストデータ作成

=> オンライン翻刻 + TEI出力

② 公開画像への校異源氏物語の頁数付与

=> くずし字OCRの活用

③ 校異源氏物語と現代語訳の対応づけ

=> TEIを用いた平行コーパスの作成

① 校異源氏物語のテキストデータ作成

①校異源氏物語のテキストデータ作成

● 作業内容

- 校異源氏物語画像のテキスト化
- OCRなどの併用

● 使用したシステム

- Omeka S + Scripto (MediaWikiを使用)

● 出力形式

- TEI/XML

Scripto · 校異源氏物語

プロジェクト デジタル源氏物語プロジェクト | 検索 Log out

校異源氏物語

Project guidelines Browse items Search item media

24 こてふ: Media #4

Original title: <https://www.dl.ndl.go.jp/api/iiif/3437687/R0000201/info.json>

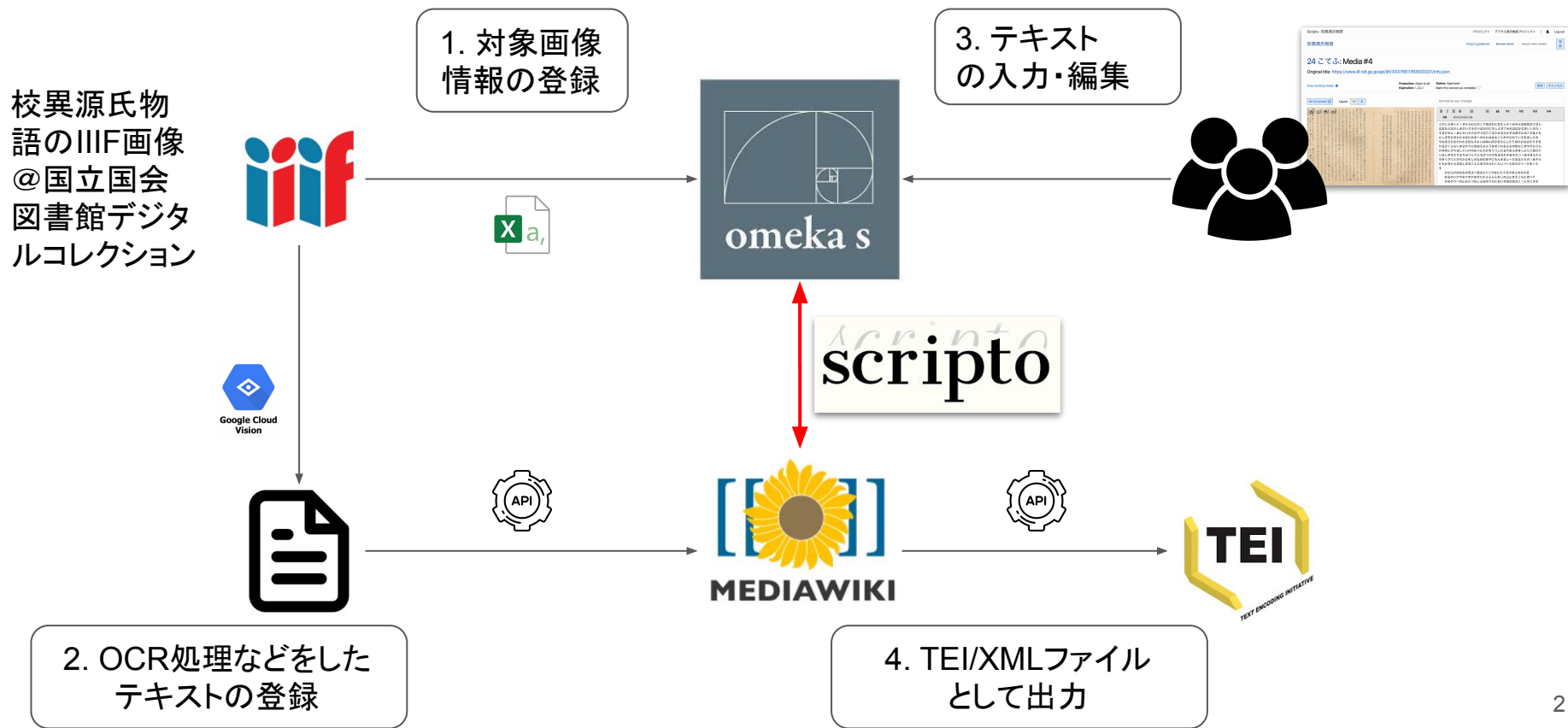
Stop tracking media ★ Protection: Open to all Expiration: しない Status: Approved Mark this revision as complete ☐ 保存 キャンセル

Go full screen Layout

Summarize your changes

B	I	U	S	≡	≡	“	h1	h2	h3	h4
h5 Horizontal rule										
とのしらぬくにききたらむ心ちしてあはれにおもしろくみならはぬ女はうなとはおもふなかしまのいりえのいはかけにさしよせてみればはかなきいのたゝすまひもたゝゑにかいたらむやうなりこなたかなたかすあひたるこそゑともしきをひきわたせるにおまへのかたははるくともみやられてをましたるやなきえたをたれたる花もえもいはぬにほひをちらしたりほかにさはかりすぎたるさくらもいまさかりにほおえみらうをめくれるふちの色もこまやかにひらけゆきにけりましていけのみにつかけをうつしたるやまふきぎしよりこぼれていみじきさかりなりみつとりものつかひをはなれすあそひつゝほそきたともをくひてとちかふをしのなみのあやにもんをまへたるなともゝゑやうにもかきとらまほしきまことにをのえもくたいつへうおもひつゝひをくら す かせふけはなみの花さへ色みえてこやなにてたてやまふきのさきはるのいけやゐてのかはせにかよふらんきしの山ふきそこにもほへりかめのうへ山のたつねしふねのうちにおいせぬなほこにのこさむ										

①校異源氏物語のテキストデータ作成(フロー)



校異源氏物語テキストDB

- 校異源氏物語テキストの TEI/XMLファイルと
行情報の RDFデータを提供するサイト
- 3つの閲覧方法を提供
 - [TEI Multi Viewer](#) (TEI-C東アジア／日本語分
科会作成)を使用したTEIテキストとIIIF画像
の並列表示
 - TEI/XMLファイル (GitHubリポジトリ)
 - [SPARQLエンドポイント](#) や、[Linked Data
Browser](#) (神崎正英氏作成)を使用したRDF
データの提供
- 利用条件: CC0



<https://kouigenjimonogatari.github.io/>



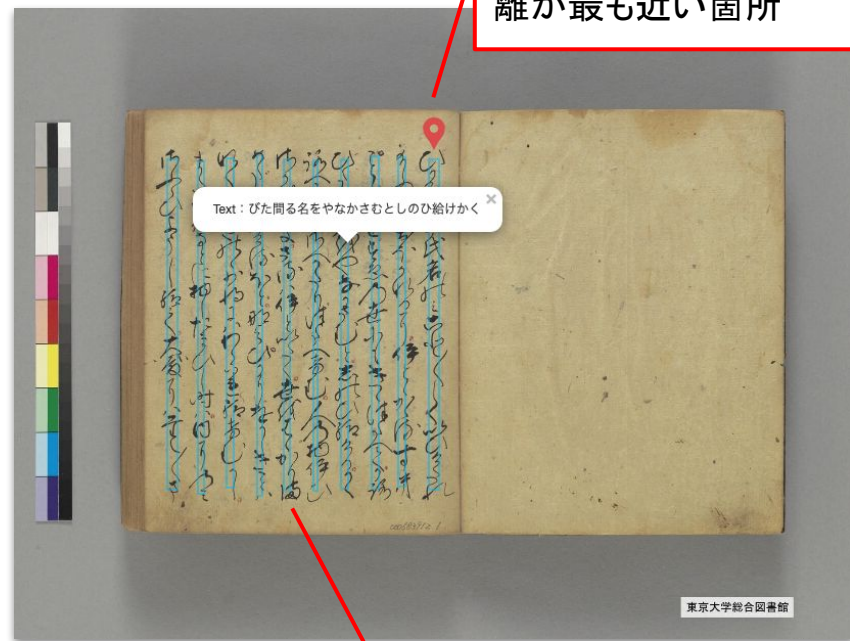
TEI Multi Viewerの表示例

② 公開画像への 校異源氏物語の頁数付与

②公開画像への校異源氏物語の页数付与

- CODHが開発しているKuroNetくずし字認識サービスを利用
 - 「自動読み順推定アルゴリズム」を利用して行単位のテキストデータを生成
- くずし字OCRテキストと、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
- 類似度が最も高い行に対して、校異源氏物語の页数を自動付与し、この結果を人手で確認
 - 概ね90%程度の精度

校異源氏物語テキストの各頁の先頭行と編集距離が最も近い箇所



くずし字OCRを使った各行のテキスト化

CODH: 人文学オープンデータ共同利用センター

日本古典籍くずし字データセット: 古典籍 44点の画像データ6,151コマから切り取った、くずし字 4,328文字種の字形データ 1,086,326文字

<http://codh.rois.ac.jp/char-shape/>

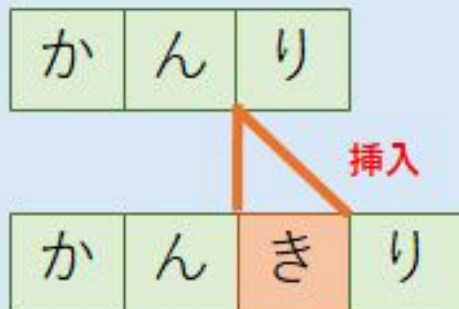
KuroNetくずし字認識サービス: IIF (International Image Interoperability Framework)に準拠した画像を対象に、多文字くずし字 OCR機能を提供

<https://mp.ex.nii.ac.jp/kuronet/>

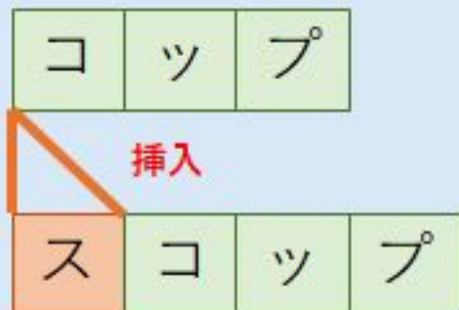


編集距離

距離1



距離1



距離3



https://mieruca-ai.com/ai/levenshtein_jaro-winkler_distance/

②公開画像への校異源氏物語の页数付与(フロー)

III対応の
公開画像



巻毎のIIFマニ
フェストファイル
を作成

Selenium
による自動化

くずし字OCR

半丁毎に画像を切り出し

KuroNetくずし字認識サービ
スにPost

くずし字OCRの実行

自動テキスト化の実行

全頁終了

No

Yes

認識結果の取得
(スクレイピング)

校異源氏物語(+ 新編日本
古典文学全集)テキストの
各頁の先頭行とOCR結果
の編集距離を算出

IIF Curation APIを用いて
頁番号の推定結果を記述:
[例](#)

人手による確認: [例](#)

IIFアノテーションリストに変
換: [例](#)

人手による確認用Googleスプレッドシート

	A	B	C	D	E	F	G	H	I
1	担当	結果	新編全集番号	新編全集テキスト	OCRテキスト	東大本ペ	URL		
166					はこしろやすくみちひきつおとこ君夢	11	http://codh.rois.ac.jp/software/iiif-4		
167		441	441	うぞおぼえたまひけんかし。女は、いと恥づかしと	かとおほえ給にもわか身いとはつかしくそ	11	http://codh.rois.ac.jp/software/iiif-4		
168					おほ給けんかし女はいとけつしと思ひしみて	11	http://codh.rois.ac.jp/software/iiif-4		
169					物し給もねひまされ御有様伊とあかぬ	11	http://codh.rois.ac.jp/software/iiif-4		
170					所なくめやすし世のたそしにも成ぬへかり	11	http://codh.rois.ac.jp/software/iiif-4		
					つる身を心もてこそかうまでもおほしゆつ	11	http://codh.rois.ac.jp/software/iiif-4		
					さるめれ衆をしり給はぬもさまなるわさ	11	http://codh.rois.ac.jp/software/iiif-4		
173					にと恨きこえ給少なすゝみいたしつるあ	12	http://codh.rois.ac.jp/software/iiif-4		
174					かきのもむきいみとろ給つやいたきぬ	12	http://codh.rois.ac.jp/software/iiif-4		
175					しか斯斯かはくちのとこそさしいへまほしか	12	http://codh.rois.ac.jp/software/iiif-4		
176					つれとの給へハ女いときくなしとおほして	12	http://codh.rois.ac.jp/software/iiif-4		
177					あさき名をいひなかしつる川くち伊かも	12	http://codh.rois.ac.jp/software/iiif-4		
178					らし関のあらかきあさましとの給さまいとこ	12	http://codh.rois.ac.jp/software/iiif-4		
179					そきたりすこしう北はらひて	12	http://codh.rois.ac.jp/software/iiif-4		
180					もりにけくきたの関を河口のあさきに	12	http://codh.rois.ac.jp/software/iiif-4		
181					のみはほせさらなむ年月のつもりもいとわり	12	http://codh.rois.ac.jp/software/iiif-4		
182		442			なくなやましきに物おほすとゑひにかち	12	http://codh.rois.ac.jp/software/iiif-4		
183			442	ず」と、酔ひにかこちて苦しげにもてなして、明く		12	http://codh.rois.ac.jp/software/iiif-4		
184					くるしけにもてなして明るもしらすかほ	12	http://codh.rois.ac.jp/software/iiif-4		
185					人くきこえわつらふを出とゝしたりかほなつ	12	http://codh.rois.ac.jp/software/iiif-4		

人手による修正結果

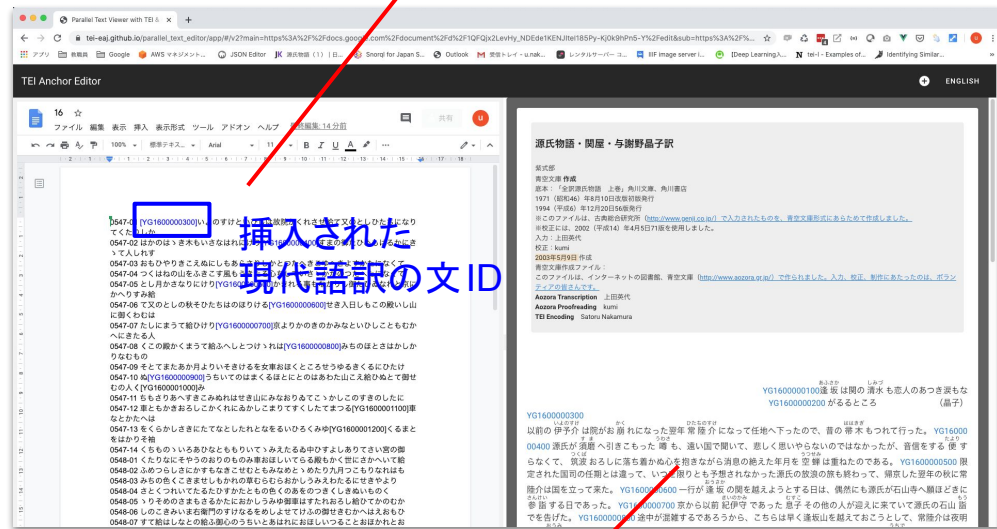
自動付与結果

③校異源氏物語と現代語訳の対応づけ

③校異源氏物語と現代語訳の対応づけ

- 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
- ①で作成した校異源氏物語の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
 - Googleドキュメントを使用して、複数人が共同で作業を実施
- 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

校異源氏物語テキストデータのGoogleドキュメント



挿入された現代語訳の文ID

現代語訳のTEI/XMLファイルをCETIcleanで表示

TEIデータの例

TEIヘッダー(例: 編集方針)

```
<encodingDesc>
  <p>
    旧字は
    <ref target="https://wwwap.hi.u-tokyo.ac.jp/ships/itaiji_list.jsp"> 史料編纂所データベース異体字同定一覧 </ref>
    を用いて新字に変換した。
  </p>
</encodingDesc>
```

```
<facsimile>
  <surfaceGrp facs="https://www.dl.ndl.go.jp/api/iiif/3437686/manifest.json">
    <surface>
      <graphic n="https://www.dl.ndl.go.jp/api/iiif/3437686/canvas/62" url="https://www.dl.ndl.go.jp/api/iiif/3437686/R0000062/full/full/0/default.jpg"/>
      <zone lrx="3445" lry="4706" ulx="0" uly="0" xml:id="zone_0085"/>
    </surface>
    <surface>
      <graphic n="https://www.dl.ndl.go.jp/api/iiif/3437686/canvas/63" url="https://www.dl.ndl.go.jp/api/iiif/3437686/R0000062/full/full/0/default.jpg"/>
      <zone lrx="6890" lry="4706" ulx="3445" uly="0" xml:id="zone_0086"/>
      <zone lrx="3445" lry="4706" ulx="0" uly="0" xml:id="zone_0087"/>
    </surface>
  </surfaceGrp>
```

IIIF画像との紐付け (半丁単位)

```
<text>
  <body>
    <p>
      <lb/>
      <pb facs="#zone_0085" n="85"/>
      <lb/>
    </p>
    <seg corresp="https://w3id.org/kouigenjimonogatari/api/items/0085-01.json">
      <anchor corresp="https://genji.dl.itc.u-tokyo.ac.jp/data/tei/yosano/03.xml#YG0300000300"/>
      ねられたまはぬま>には我はかく人にくまれてもならはぬをこよひなむはし
    </seg>
  </body>
</text>
```

校異源氏物語と現代語訳の対応づけ

まとめ：構築したデータ(主な作業協力者：5名)

	方法	作成したデータ	数量
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	54巻中の54巻 1,812頁856,495文字
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	3,977頁1,687,192文字
	- 編集距離 - 人手による確認	IIIF Curation API準拠のJSONファイル	- 東大本 - 九大本(2種類) - 湖月抄(54巻中の23巻)
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	54巻中の9巻

結論

考察

- くずし字OCRと編集距離を用いた校異源氏物語の頁番号の自動付与
 - 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が可能となった => 作業コストの削減 => Ver.YUMENOUKIHASHIの公開
 - 源氏物語研究の専門家から、くずし字 OCRそのものの有用性に加えて、OCR結果の不完全さを補う手法としての編集距離が高く評価された。

結論

- 『デジタル源氏物語』システムにおけるデータ構築について述べた。
 - 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

今後の展望

- 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
- 頁数の推論結果に対する人手によるチェック
- IIIFキュレーションリストへの変換

協力者(第三者・各所蔵機関)にご協力いただきたい項目

- 対象

- 源氏物語
- その他の古典籍

Home

画像とテキストを一緒にみる

▶ 頁数で探す

校異源氏物語

新編日本古典文学全集

≡ IIIF対応源氏物語リスト

⚠ 東大本・脱文錯簡リスト

ℹ このサイトについて

📧 お問い合わせ

デジタル源氏物語

🌐 日本語 ▼

お問い合わせ

データ連携や作業協力を検討されている方は、下記の問い合わせ先までご連絡をお願いいたします。
お問合せの内容によっては回答できない場合や時間を要する場合がありますこと、また、ご要望に沿えない場合もあることを、あらかじめご了承ください。
入力していただいた個人情報は、今回の問合せ手続のみに使用し、他の目的には使用しません。

お問い合わせ先

digital-genji-public@googlegroups.com

36

謝辞

本システム開発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。

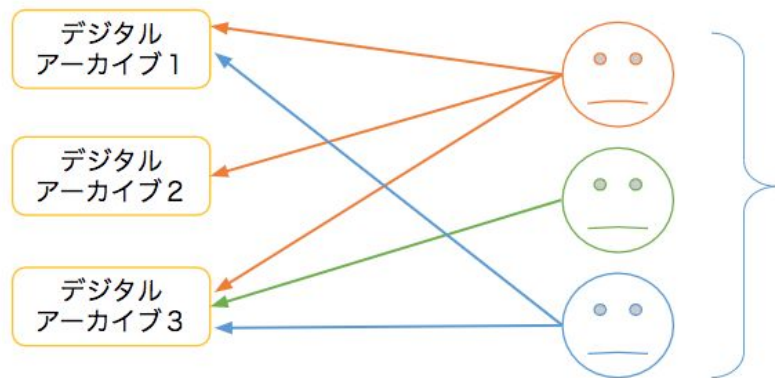
またくずし字OCRの利用にあたっては、CODH北本朝展先生、カラーヌワット・タリン先生にご協力いただきました。深く感謝いたします。

本研究はJSPS科研費 [19K20626](#)の助成を受けたものです。

ご清聴ありがとうございました。

その他

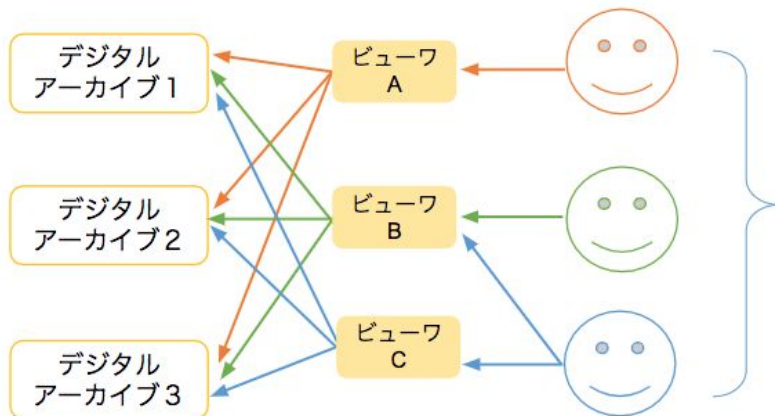
従来のデジタルアーカイブ: 同じようなものが別々に作られ、そこでしか見られない状態



- それぞれ使い方を覚えないと..
- それぞれアクセスしないと..
- それぞれ検索しないと..
- それぞれの利用条件を確認しないと..
(でも、よくわからない...)

※提供側:「公開しても、なかなか
見つけてもらえない」

IIIF対応デジタルアーカイブ: 相互運用可能、共通インタフェース、充実した機能の提供



- 一つのビューワを通して、複数機関の画像を見ることができる。(IIIF対応ビューワの種類は複数あり。)
- ビューワが各デジタルアーカイブとやりとりしてくれるので、利用者の学習コスト・探索の手間が大幅に低減。
- ライセンス表示も共通化されるので、利用条件が把握しやすい。
- 高精細画像でもスムーズな表示が可能。

※提供側のメリット: 画像共有機能を持つことで、発見可能性が高まる。

結果

- 詳細
 - [東大本](#)
 - [湖月抄\(国文研\)](#)
- 東大本:96.8%
- 湖月抄(国文研):84.7%

※ ただし、一行のズレは許容(厳密には先頭行を指せていないケースあり)