
古典籍データ駆動研究センターの目指すもの GOALS OF THE CENTER FOR DATA- DRIVEN RESEARCH ON PREMODERN JAPANESE TEXTS

菊池信彦（古典籍データ駆動研究センター特任准教授）

Nobuhiko Kikuchi, Ph.D (Specially Appointed Associate Professor, the Center for Data-Driven Research on Premodern Japanese Texts)



“Ukiyoe-style painting depicting a man programming in front of a desktop” generated by Stable Diffusion.

はじめに

Introduction

古典籍データ駆動研究センターとは

What is the Center for Data-Driven Research on Premodern Japanese Texts?

古典籍データ駆動研究センターとは

- 人文学分野の研究者と情報科学・自然科学分野の研究者との協働によって、
- 人文学研究の新たな領域を切り拓くデータ駆動型研究や異分野融合研究を推進し
- 他分野と協働し得る課題解決型の人文学研究の創出を目指す組織

The Center for Data-Driven Research on Premodern Japanese Texts is...

an organization that promotes data-driven research and interdisciplinary research that opens up new areas of humanities research through collaboration between researchers in the humanities and those in the information and natural sciences, and aims to create problem-solving humanities research that can collaborate with other fields.

センターの目的 / The aims of our center

- 歴史的典籍ネットワーク事業の成果を基盤とした、「データ駆動型」人文学研究の推進
 - 人文学分野と自然科学・社会科学分野との協働によるデータの利活用を通して現代社会の課題解決に資する新たな研究領域、すなわち「総合知」の創出に貢献
 - 人間文化研究機構内の他の研究機関のDH関係の研究者らとの協働で、機構によるDHプロジェクトの推進にも参画
- Promotion of "data-driven" research in the humanities based on the results of the NIJL-NW Project
 - Contribute to the creation of new research fields, or "comprehensive knowledge," that will contribute to solving issues in contemporary society through the utilization of data in collaboration with the humanities and the natural and social sciences.
 - Our center also participates in the DH projects by the National Institutes for the Humanities(NIHU) in collaboration with DH-related researchers from other institutes within the NIHU.
-

本報告の目的 / The Purpose of this Presentation

弊センターのデータ駆動型研究領域について、特に報告者が進めている研究プロジェクトを取り上げる

In my presentation, I will introduce the data-driven research areas of our center, with a particular focus on the research projects I am working on.

データ駆動型 研究プロジェクト群

the Data-
Driven
Research
Areas of Our
Center

研究基盤整備部門／ Research Infrastructure Development Division

- 古典籍デジタル画像データの機械可読化のためのシステムを整備するとともに、そこで生成された機械可読化データの蓄積と構造化及び公開を行う「データインフラストラクチャの設計と構築」
- 蓄積されたデータに基づく「人文系データ分析技術の開発」
→ 本日も話題になっているくずし字OCR技術を活用した本文テキストデータの作成と、それをより機械的に分析できるようにすべくTEIマークアップの研究等。これについては、後ほど詳しくます。

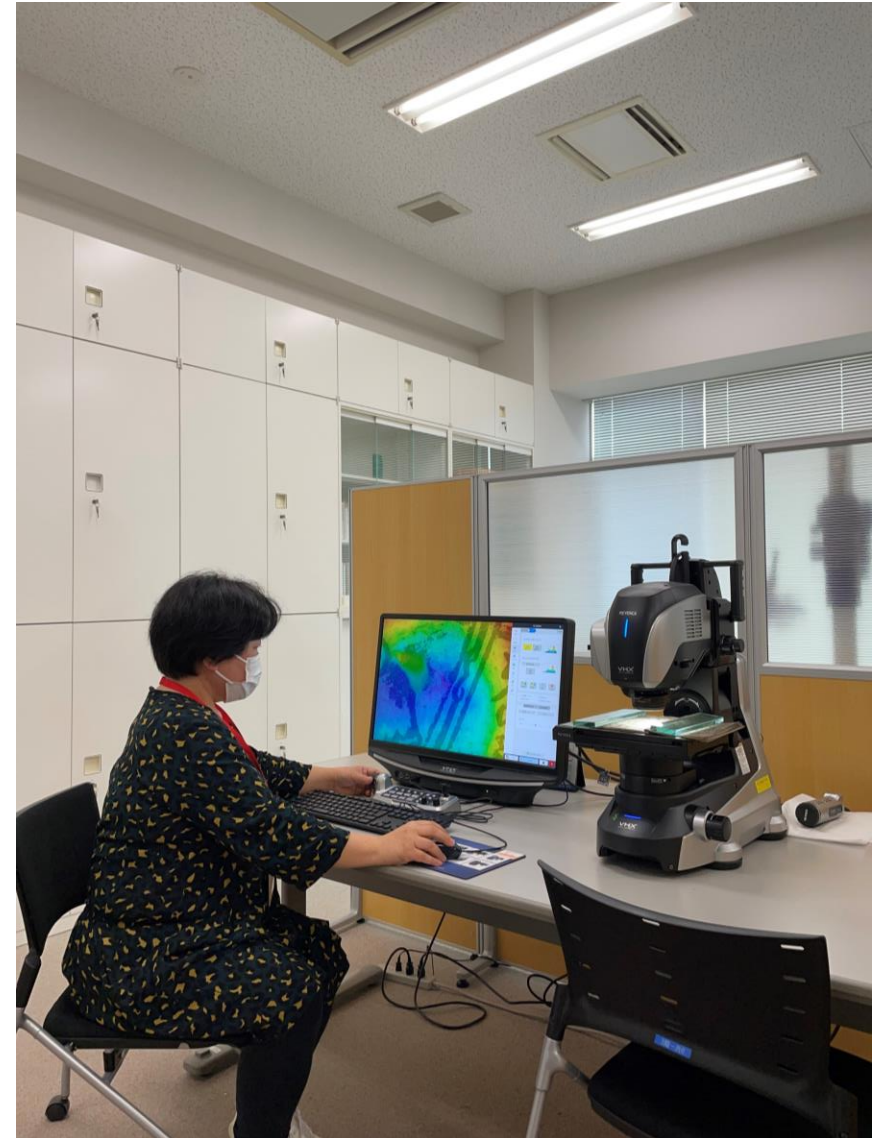
- The Research Infrastructure Development Division is in charge of "design and construction of data infrastructure" to develop a system for machine readability of digital image data of classical manuscripts and to store, structure, and publish the generated machine readable data, as well as "development of humanities data analysis technology" based on the stored data.
→ Specifically, I am researching the creation of textual data of classical manuscripts and TEI markup to enable more mechanical analysis of such data, utilizing the kuzushiji OCR technology that has been the topic of discussion today. I will discuss this in more detail later.
-

データ駆動型研究推進部門／ Data-driven Research Promotion Division

- 蓄積されたデータコンテンツを十全に活用し、他分野との協働により現代社会の課題解決に資する研究を行う「コンテンツ解析」
 - 物質として書物が蓄積してきた素材情報（構成要素としての紙の情報、混入物としての製作者・読者の毛髪や皮脂等の人間由来の情報）の分析とデータの蓄積を行う「マテリアル分析・解析」を担当
 - The Data-driven Research Promotion Division is in charge of "Contents Analysis," which conducts research that contributes to solving issues in modern society by fully utilizing accumulated data contents and collaborating with other fields.
 - This division is also in charge of "Material Analysis," which analyzes and accumulates data on material information that books have archived as substances (information on paper as a component and information of human origin such as hair and sebum of the creator and readers as a mixture).
-

マテリアル分析の様子 Material analysis research in progress...

- 松原哲子（機関研究員）氏が高精細デジタルマイクロスコープを使い、モノ史料としての草双紙の夾雑物等の調査研究を行っています。
- Dr. Noriko Matsubara (Institutional Fellow) is using a high-resolution digital microscope to conduct research on impurities and other materials in Kusa-Zoshi as physical materials.



データインフラおよび
人文系データ分析技術としての
廣瀬本万葉集TEI化プロジェクト

Hirose-bon
Man'yoshu TEI
project as data
infrastructure and
humanities data
analysis
technology
research

プロジェクト概要／General Overview of the Project

- 廣瀬本万葉集とは「非仙覚本系統」に属する唯一の完本として、万葉集伝本研究上重要な位置を占める資料
 - プロジェクトの目的：廣瀬本万葉集をTEI/XML化→データ駆動型研究の実践に向けた礎に
 - メンバー：菊池信彦（国文研）、海野圭介（国文研）、乾善彦（関西大）、永崎研宣（人文情報学研究所）、吉賀夏子（佐賀大）ほか
 - The Hirose-bon Man'yōshū is the only complete manuscript belonging to the "non-Senkaku lineage" and is an important source for the study of Man'yōshū manuscripts.
 - Project Objective: To convert the Hirose-bon Man'yōshū text to TEI/XML as a basis for data-driven research in humanities
 - Members: Nobuhiko KIKUCHI (ILCAA), Keisuke UNNO (ILCAA), Yoshihiko INUI (Kansai Univ.), Kennobu NAGASAKI (NIIHS), Natsuko YOSHIGA (Saga Univ.) and others
-

廣瀬本万葉集TEIマークアップの方針

Hirose-bon Man'yoshu TEI Markup Policies

- ①訓がひらがなかカタカナか、付訓の位置が歌本文の右か左か、またその変化はどうか、題詞のインデントが歌と比べて高いか低いか（ Whether the kun is in hiragana or katakana, whether the supplementary kun is to the right or left side of the verse text and how it changes, and whether the indentation of the title is higher or lower than that of the verse. ）
- ②墨字の注・割書・合点のなかで藤原定家由来とするものを判断し、タグに記入すること（To determine which of the notes and marks in black ink are attributed to Fujiwara no Teika, and enter them in a tags）

上記の方針は、万葉集伝本研究において着目すべき情報を取り出すことに向けられています
(The above policies aim to extract information that should be focused on in the study of the Man'yoshu manuscripts)

```

742 </div>
743 <div xml:id="旧国歌大観番号一三七">
744 <lg xml:id="many0137" type="短歌">
745 <l xml:id="many0137_01"><lb/>秋山尔落黄葉須與者勿散乱曾妹之當将見<note type="割書" resp="#万葉集">一云知里勿<milestone unit="br"/>乱曾</note></l>
746 <l xml:id="manyokun0137_01" corresp="#many0137_01"><lb/>アキヤマニオツルモミチハシハラク
747 <note type="合点" subtype="異訓" resp="#定家" place="right">ゝハ</note>モ<anchor type="noteEnd"/>チリナミタレソ
748 <note type="注" resp="#定家" place="right">伊云御本云イモ、ミルヘク同本云御本云<milestone unit="br"/>如今本</note>イモカアタリミ
749 </lg>
750 </div>
751 <div xml:id="旧国歌大観番号一三八">
752 <p style="text-indent:3em">或本歌一首<note type="割書" resp="#万葉集">并短哥</note></p>
753 <ab xml:id="many0138" type="長歌">
754 <lb/><seg type="ruby"><seg type="rb">石見之海津乃浦乎無美浦無跡人社見良米瀉無跡人</seg>
755 <lb/><seg type="rt" rend="right">イハミノウミつのうらをなみうらなしと人こそみらめかたなしと人</seg></seg>
756 <pb n="六一丁表" facs="https://www.iiif.ku-orcas.kansai-u.ac.jp/iiif/2/210464810%2F0059.tif/full/full/0/default.jpg"/>
757 <lb/><seg type="ruby"><seg type="rb">社見良目者咲八師浦者雖無縦恵夜思瀉者雖無勇魚取</seg>
758 <lb/><seg type="rt" rend="right">こそみらめよしゑやうらはなけれとよしゑやしかたはなけれとくちらとる</seg></seg>
759 <lb/><seg type="ruby"><seg type="rb">海邊乎指而柔田津乃荒磯之上尔
760 <note type="注" resp="#定家" place="right">伊云敷同本云御本云蚊</note>蚊<anchor type="noteEnd"/>青生玉藻息都藻</seg>
761 <lb/><seg type="rt" rend="right"><note type="異訓" resp="#定家" place="right">アマ</note>うみ<anchor type="noteEnd"/>
762 へをさしてにきたつのあらいそのうへにかあをゝふるたまもおきつも</seg></seg>
763 <lb/><seg type="ruby"><seg type="rb">明来者浪己曾来依夕去者風己曾来依浪之共波依玉藻</seg>
764 <lb/><seg type="rt" rend="right">あけくれはなみこそきよれゆふされはかせこそきよれなみのともなみよるたまも</seg></seg>
765 <lb/><seg type="ruby"><seg type="rb">成靡吾宿之敷妙之妹之手平露霜乃置而之来者</seg>
766 <lb/><seg type="rt" rend="right">なりなひきわれかやとりしきたへのいもかたもとをゆしものおきてゝくれは</seg></seg>
767 <lb/><seg type="ruby"><seg type="rb">此道之八十隈毎萬段顧雖為弥遠尔里放来奴益高尔</seg>
768 <lb/><seg type="rt" rend="right">このみちのやそくまことによろつたひかへりみすれというやと
769 <note type="異訓" resp="#定家" place="right">ハ</note>を<anchor type="noteEnd"/>にさとはなれきぬいやたかに</seg></seg>
770 <lb/><seg type="ruby"><seg type="rb">山毛越来奴早敷屋師吾嬢乃児我夏草乃思志姜而</seg>
771 <lb/><seg type="rt" rend="right">山もこえきぬ<note type="異訓" resp="#定家" place="right">ハ</note>はしきやし</note>
772 さしきやのし<anchor type="noteEnd"/>わかづまのこかなつくさのおもひしなえて</seg></seg>
773 <lb/><seg type="ruby"><seg type="rb">将嘆角里将見靡此山</seg>
774 <lb/><seg type="rt" rend="right">なけゝとものゝさとみむなひけこの山</seg></seg>
775 </ab>

```

現在のところ、旧国歌大観番号85～140まで作成
As of now, we have created the old Kokka Taikan numbers 85-140.

TEI/XMLデータは
[GitHub](#)で公開
(CC-BY)
TEI/XML data is available on
[GitHub](#) (CC-BY).

タグの分析を通じた万葉集伝本研究への貢献

Contribution to the Manyoshu studies through the analysis of TEI tags

タグの分析を通じて、どこにどのような書入れがあるか（図参照）、属性として記述した「#定家」の情報をもとに定家由来の書入れがどこにどれほどあるのか一覧／分析できる。

Through the analysis of tags, users can see where and what kind of calligraphy is stored (see figure), and list/analyze where and how much calligraphy of Teika origin is stored based on the information of "#定家" described as an attribute.

@place 属性の値	件数
right	152
none	32
left	22
margin-head	6
margin-right	2
head	1
計	215

書き入れにおける@place 属性の値の登場件数

出典：永崎ほか, 2021.

廣瀬本万葉集TEI/XMLビューワの開発

Development of a TEI/XML Viewer for the Hirose-bon Man'yoshu

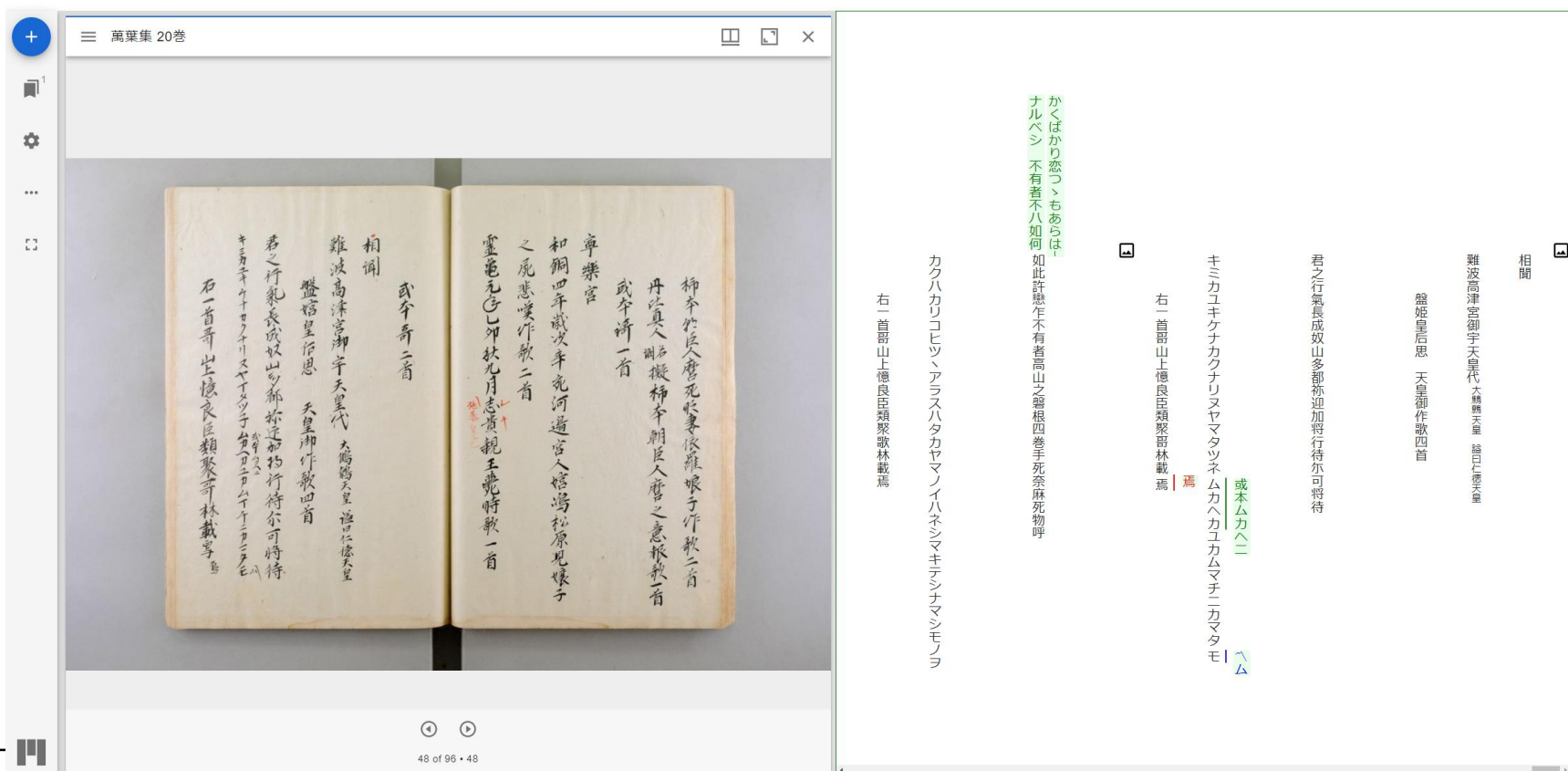
皆が皆XMLデータの解析に慣れているわけではないので、それを手軽に行うことのできるビューワも開発（永崎）

Not everyone is familiar with parsing XML data, so we also developed a viewer that makes it easy to do so. (it has been developed by Dr. Nagasaki)

<https://candra.dhii.jp/nagasaki/manyo/manyoviewer2021.html>

廣瀬本『万葉集』 TEI/XMLビューワ

検索モード／画像閲覧モード



データ駆動型人文学の創生に向けて

Toward the Creation of a Data-Driven Humanities

- データ駆動型人文学のための環境を作ること
 - ①和歌に焦点をあてて、他の和歌集のTEIマークアップのモデル
 - ②「ルビ」と「訓」のマークアップに関する検討
 - ③①と②を踏まえ（和歌集を含む）古典籍資料マークアップの実践モデル
- データ駆動型人文学の環境を使った実践を行うこと
 - ④TEIタグの統計的・情報学的分析

Our goal is to create an environment for data-driven research in humanities.

① Focusing on waka poetry, our goal is to create a model for TEI markup of other waka poetry collections.

② We will contribute to the international discussion by conducting a series of studies on “rubi” and “kun” markups.

③Based on the both above, we aim to create a practical model for markup of classical materials (including waka poetry anthologies) and use it as an educational resource.

We are also considering to practice using a data-driven research in humanities environment.

④ By conducting statistical and informatics analysis of TEI tags, we aim to analyze waka poetry in a way that can only be done by data analysis.

おわりに

Conclusion

弊センターは
他分野との協
働を求めています

Our center
seeks
opportunities
of the
collaboration
with other
disciplines
inside/outside Japan



“Ukiyoe depicting many researchers getting together in a friendly atmosphere” generated by Stable Diffusion.

- 古典籍データ駆動研究センターは、データ駆動型の人文学、いわゆるDHの推進と、そしてそれが本質的にもつ他分野との協働という特徴があります。
Our center is characterized by its promotion of data-driven humanities, or the DH, and collaboration with other disciplines that the DH essentially has.
- 国内外の研究者・研究者コミュニティとの連携を常に求めていますので、古典籍を活用する研究領域のかた、当館データで何か新しいものを求める方は、どうぞお気軽にご連絡ください。(We are always looking for collaboration with domestic and international researchers and research communities, so please feel free to contact us if you are in a research field that utilizes premodern Japanese books or if you are looking for something new with our data.)