Vol.2023-CH-131 No.8
2023/2/18

# 和歌の XML/TEI データ分析のための自主学習環境の構築

菊池信彦[1]　永崎研宣[2]　乾善彦[3]　海野圭介[4]　小川歩美[5]　吉賀夏子[6]

**概要**：概要：近年，デジタルヒューマニティーズの浸透とその興隆を受け，人文学においても多種多様な研究データがオープンデータとして公開されるようになっている．しかし，オープン化されたとはいえ，だれもがそれらの扱いに長けているわけではない．この状況を受け，オープンデータを用いて自ら学習できる環境やツールの整備もまた，近年急速に広まりつつある．報告者らは，構造的なテキストデータとして取り扱いやすい和歌を教材とした自主学習サイトの構築を目指すこととした．本発表ではその事例の一つとして，国文学研究資料館および関西大学との連携に基づいて作成を進めている廣瀬本万葉集 TEI/XML データを採りあげ，その途中経過について報告する．

**キーワード**：和歌，TEI/XML，自主学習リソース，廣瀬本万葉集，古典籍，データ駆動型人文学

# Creation of a Self-Learning Environment for Analysis for TEI/XML Data of Waka Poetry

NOBUHIKO KIKUCHI[†1]　KIYONORI NAGASAKI[†2]
YOSHIHIKO INUI[†3]　KEISUKE UNNO[†4]　AYUMI OGAWA[†5]　NATSUKO YOSHIGA[†6]

***Abstract***: In recent years, with the spread and rise of digital humanities (DH), a wide variety of research data has become available as open data in the humanities. It has been accessible, however, it is not easy to handle for everyone. In response to this circumstance, the development of environments and tools for self-learning using research open data has also been spreading rapidly in recent years. We decided to construct a self-learning environment using Waka poetry, which is easy to handle as structured text data, as a learning material. In this presentation, we will report on the progress of the development of the environment for analyzing the TEI/XML data of the Hirose-Bon Man'yoshu, which is being created in collaboration with the National Institute of Japanese Literature (NIJL) and Kansai University, as an example.

***Keywords***: Waka Poetry, TEI/XML, Self-Learning Resources, Hirose-Bon Man'yoshu, Premodern Japanese Texts, Data-Driven Humanities

## 1. はじめに

近年，世界的なデジタルヒューマニティーズ（以下，DH）の浸透とその興隆を受け，人文学においても多種多様な研究データがオープンデータとして公開されるようになっている[1]．しかし，オープン化されたとはいえ，だれもがそれらの扱いに長けているわけではない．それはとりわけデジタル技術から縁遠いとみなされている人文学研究者にとってはより顕著であり，また，大きな課題ともなっている．

この状況を受け，研究用オープンデータを用いて自ら学習できる環境やツールの整備もまた，近年急速に広まりつつある．そこで報告者らは，構造的なテキストデータとして取り扱いやすい和歌を教材とした自主学習サイトの構築を目指すこととした．本報告ではその事例の一つとして，国文学研究資料館および関西大学との連携に基づいて作成を進めている廣瀬本万葉集 TEI/XML データ[2][3]を採りあげ，その途中経過について報告する．

1 国文学研究資料館
National Institute of Japanese Literature
2 人文情報学研究所
International Institute for Digital Humanities
3 関西大学
Kansai University
4 国文学研究資料館
National Institute of Japanese Literature
5 合同会社 AMANE
AMANE.LLC
6 佐賀大学
Saga University

## 2. 先行研究・事例の整理とその課題

筆者の一人である菊池は国内外の DH の教育・学習リソースの状況について論じたことがある．そこでは，DH の教育・学習リソースは欧米で盛んであり，それに比べると日本および東アジア研究のための DH 研究のノウハウはまだ十分に広まっておらず，とりわけ，オンライン環境の開発が課題であることを指摘した[4].

DH の普及と教育のための環境の開発は様々な研究者・教育研究機関により多方面から継続的な取組みが進められており，日本語での DH 研究の学習環境は徐々に改善されつつある．しかし，日本語文化圏全体としてみた場合，全体としてはまだ不十分であり，日本文学研究に関しても多くの分野がカバーされていない．さらに最近では，国立国会図書館デジタルコレクションで OCR テキストデータが

公開され，大規模テキストの大雑把な検索が可能となったものの，構造化されたテキストではなく，正確性も十分ではないため，既存の人文学的な研究手法と整合させることについては大きな課題となっている．

人文・社会科学のための研究データインフラストラクチャーの構築と提供を目指す，日本学術振興会による人文学・社会科学データインフラストラクチャー構築推進事業[5]では，人文・社会科学のための研究データの活用環境を提供すべくオンライン分析ツールを提供しているが，これはJupyter Notebook環境の提供が行われているのみであり，具体的にどのように活用するかについては主に研究者コミュニティに委ねられている．

そこで，報告者らは，日本文学研究に関する分野のうち，特に和歌資料を対象にした今後のTEI/XMLデータ作成と共有の興隆を見据え[a]，近い将来必要となるデータ分析のための研究ノウハウ共有に向けて，自主学習のための環境を構築することとした．

## 3. DH学習環境「廣瀬本万葉集 TEI/XMLデータ分析入門」の構築

### 3.1 構築方針

報告者らは構築にあたり以下を方針と定めた．

一点目は汎用性と専門性を両立させること．すなわち専門的な資料―本報告においては廣瀬本万葉集を指す―の分析を進めながら，同時に今後登場するであろう他のTEI/XMLデータ分析にも応用できるような知識習得を目指した．

二点目は「手を動かす」こと．テキストを読むあるいはレッスン動画を見て終わりにするのではなく，実際にコーディングをしながら，あるいは，プログラムを実行することで，その成果が目に見えるように配慮した．また，ユーザ自身が求める内容を得られるようにコード内容をカスタマイズしやすくすべく，丁寧なコメントを心掛けた．

三点目は学習・実践環境の構築に負担が少ないこと．各種ソフトウェアのインストールをはじめとする学習環境を整えるための苦労は，それだけで導入へのハードルを上げ，学習意欲を削ぐことになる．したがって，サイトを開いて即座に実践できる環境を選択することとした．

以上のことから，先行事例でも利用され，また，後二者の条件を満たす環境として，ここでは，Google アカウントさえ有していれば，あとはWebサイトにアクセスするだけでPythonの実行環境の利用が可能なGoogle Colaboratoryを採用した[10]．

### 3.2 レッスンの構造とその内容

作成した「廣瀬本万葉集 TEI/XML データ分析入門」は次の図1の通りである．


図 1　開発中の「廣瀬本万葉集 TEI/XML データ分析入門」（Google Colaboratory）

レッスンは，「大きな」情報からはじめ，ユーザ自身が必要とする「小さな」情報を抽出する方法へと，段階を追って進む構成とした．具体的には，TEIの紹介から始まり，TEIファイル全体やテキスト全体を抽出する方法を学び，その後，特定の要素や属性に焦点を当てて検索，その結果を表示あるいはダウンロードする方法を学ぶという構成である．これは，レッスンを作成するにあたって参考にしたCuper[11]と永崎の事例に倣ったものである[12]．

しかし，レッスン構成としては他の事例を参照しつつも，その内容は，廣瀬本万葉集特有の事情を反映することで，本資料を扱う研究者の関心に沿った情報抽出ができるように配慮した．すなわち，廣瀬本万葉集は，和歌テキストの内容だけでなく，写本として記載されている様々な付加情報に研究上の価値が見出され，すでに研究成果として公表されている資料である[13]．そのため，報告者らは，ここで利用するTEI/XMLデータに対しては，この「特有の事情」を反映する形で，内容よりも付加情報の様態に焦点をあてたマークアップを行っているところである．そして，これを踏まえ，レッスン内容は，先述の「小さな」情報と表現した，歌や本文の種類ごと（短歌，長歌，それ以外の本文），付与されている情報（和歌検索のための索引番号である「旧国歌大観番号」や人物名）ごと，そして本文以外に付与された「書き入れ」の由来や位置情報などを抽出できるようなものとした．

上述の通り，作成したレッスンは，TEIファイルからの情報の検索と抽出，そしてテキストファイルでのダウンロードに焦点を当てたものである．加えて，レッスン最後にはグラフ化に関する教材も掲載することで，次のステップ

a) 古典籍一般としたTEI/XMLマークアップに関しては，岡田が，国文学研究資料館による「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」研究開発系共同研究「TEIの導入」において，TEIコンソーシアム東アジア・日本語分科会の協力のもと，そのガイドラインの作成を試みている[6]．
2021年には，TEIガイドラインに準拠したテキストデータ構築をテーマにした入門書も刊行されており[7]，今後日本語資料を対象にしたTEIの普及が期待される．
また，現時点ではTEI/XML形式のデータではないものの，山元啓史らは八代集テキストデータセットおよび語彙データセットをオープンデータとして公開しており[8][9]，和歌テキストのオープンデータとして様々な活用可能性が期待される．

である可視化分析への導入とした[b].

## 4. おわりに：課題と今後に向けて

「廣瀬本万葉集 TEI/XML データ分析入門」は，それが対象とする廣瀬本万葉集 TEI/XML データともども，作成途上のものである．今後は，データ自体の拡充とともに，広く公開しユーザからのフィードバックを受けながら，レッスン内容の改訂・追加を重ねつつ，他の和歌集や古典籍データへの応用可能性を模索し，充実化を図っていきたい

## 参考文献

[1] The Journal of Open Humanities Data (JOHD) https://openhumanitiesdata.metajnl.com/, (accessed 2023-01-24)
[2] 永崎研宣, 乾善彦, 菊池信彦, 宮川創, 小川歩美, 堀井洋, 吉賀夏子. 万葉集伝本研究のためのデジタル基盤構築：廣瀬本『万葉集』の構造化とビューワの開発. 研究報告人文科学とコンピュータ（CH）. 2021, 2021-CH-125(2), pp.1-7.
[3]『廣瀬本万葉集』翻刻＆TEI 化プロジェクト. https://github.com/KU-ORCAS/manyoshuTEI, (参照 2023-01-25)
[4] 菊池信彦, 宮川創, 二ノ宮聡. 「東アジア DH ポータル」の構築と課題：デジタルヒューマニティーズの研究ノウハウのオープンな知識基盤を目指して. じんもんこん 2020 論文集. 2020, pp.229-234.
[5] 人文学･社会科学データインフラストラクチャー構築推進事業. https://www.jsps.go.jp/j-di/torikumi.html, (参照 2023-01-26)
[6] 岡田一祐. 日本古典籍テクストの TEI/XML による符号化ガイドライン作成のこころみ. https://www.academia.edu/42260399/日本古典籍テクストの TEI_XML による符号化ガイドライン作成のこころみ. （参照 2023-01-25)
[7] 一般財団法人人文情報学研究所（監修）. 石田友梨, 大向一輝, 小風綾乃, 永崎研宣, 宮川創, 渡邉要一郎 編. 人文学のためのテキストデータ構築入門 : TEI ガイドラインに準拠した取り組みにむけて. 文学通信, 2022.
[8] Yamamoto, Hilofumi, Hodošček, Bor. Hachidaishu part of speech dataset (1.0.0) [Data set]. Zenodo. 2021. https://doi.org/10.5281/zenodo.4835806, (accessed 2023-01-26)
[9] Yamamoto, Hilofumi, Hodošček, Bor. Hachidaishu vocabulary dataset (1.0.1) [Data set]. Zenodo. 2021. https://doi.org/10.5281/zenodo.4744170, (accessed 2023-01-26)
[10] “廣瀬本万葉集 TEI/XML データ分析入門“. https://colab.research.google.com/drive/1L-olwb1OBx_SYiZ9xaf94nbfo9OB5AVU?usp=sharing, (参照 2023-01-21).
[11] Cuper, M., Boer, E. den, Automatically extract XML content with Python. KB Lab: The Hague. 2022. https://lab.kb.nl/tool/automatically-extract-xml-content-python, (accessed 2023-01-21)
[12] 永崎研宣. 「人文学のためのテキストデータ構築入門」フォローアップサイト. 人文情報学研究所. https://www.dhii.jp/dh/tei/, (参照 2023-01-21)
[13] 田中大士. 衝撃の「万葉集」伝本出現 : 廣瀬本で伝本研究はこう変わった. 塙書房, 2020 年.

b) なお，可視化に関しては，著者の一人である永崎がすでに廣瀬本万葉集のためのビューワを開発している．参考文献[2]を参照のこと．