

# 古典籍画像データを利用したTEI/XMLデータ作成 —廣瀬本万葉集を中心に—

菊池信彦（研究代表／国文学研究資料館／kikuchi.nobuhiko@nijl.ac.jp）

## 1. 本研究プロジェクトの背景

「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」（2014-2023年度）

国文学研究資料館は、国内外の大学等の研究機関・図書館と連携し、日本語の歴史的典籍約30万点全冊のデジタル画像化を行い、「国書データベース」として公開した。

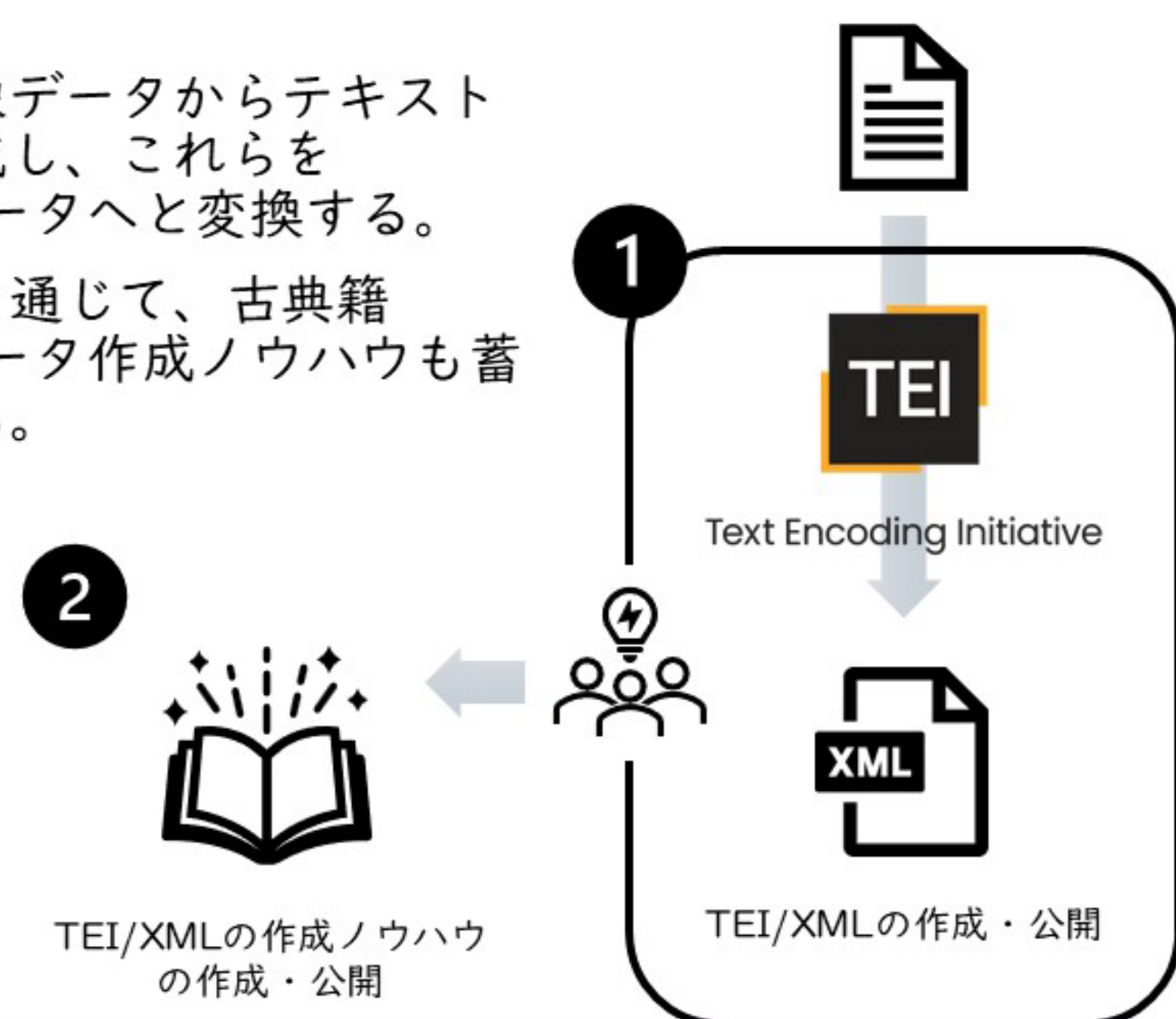
「データ駆動による課題解決型人文学の創成～データ基盤の構築・活用による次世代型人文学研究の開拓～」（2024年度-2033年度予定）

くずし字OCR技術向上を背景に、画像からテキストデータの作成が進め、そのビッグデータを基にした国文学、人文学、そして「総合知」の創出を目指す



## 2. プロジェクトの目的

- ①古典籍画像データからテキストデータを作成し、これらをTEI/XMLデータへと変換する。
- ②研究活動を通じて、古典籍TEI/XMLデータ作成ノウハウも蓄積、公開する。



国文学研究者に限らず、国内外の古典籍所蔵機関および異分野の研究者コミュニティに発信し、もってデータ駆動型人文学のための研究資源の確立を目指す。

## 3. TEIとその意義

- TEI (Text Encoding Initiatives) とは？

人文系テキストデータのXMLマークアップガイドラインであり、それを策定する団体。人文学研究資料の構造化データとして、TEIはすでに国際的なデファクトスタンダードとなっている。

国内では「TEI-C東アジア/日本語分科会」の発足を背景に、普及が徐々に進みつつある状況。

- TEI/XMLデータ作成の意義  
日本の古典籍資料データが国際的な研究の場での流通に耐えうる形となる。その結果、比較文学研究等のグローバルな視座での研究にも役立つ研究資源になりうる。



## 4. TEI/XMLデータ化対象資料

- メンバーには、国文研内教職員（国文学、人文情報学）を中心に、館外からも参加
- TEI/XMLデータ化対象は、メンバーそれぞれの研究関心・対象に応じて様々。
  - ①古典籍一般の書誌情報：国書データベースの書誌データ
  - ②古代・中世期の和歌資料：廣瀬本万葉集、古今和歌集
  - ③近世期の貸本屋関連資料と漂流記、旅行記 等

以下、菊池が主担当で進めている廣瀬本万葉集のプロジェクトを紹介

## 5. 廣瀬本万葉集のTEI/XMLマークアップ

- 廣瀬本万葉集：「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。本研究は所蔵者の関西大学と協定を結び実施。

- 廣瀬本万葉集プロジェクトメンバー

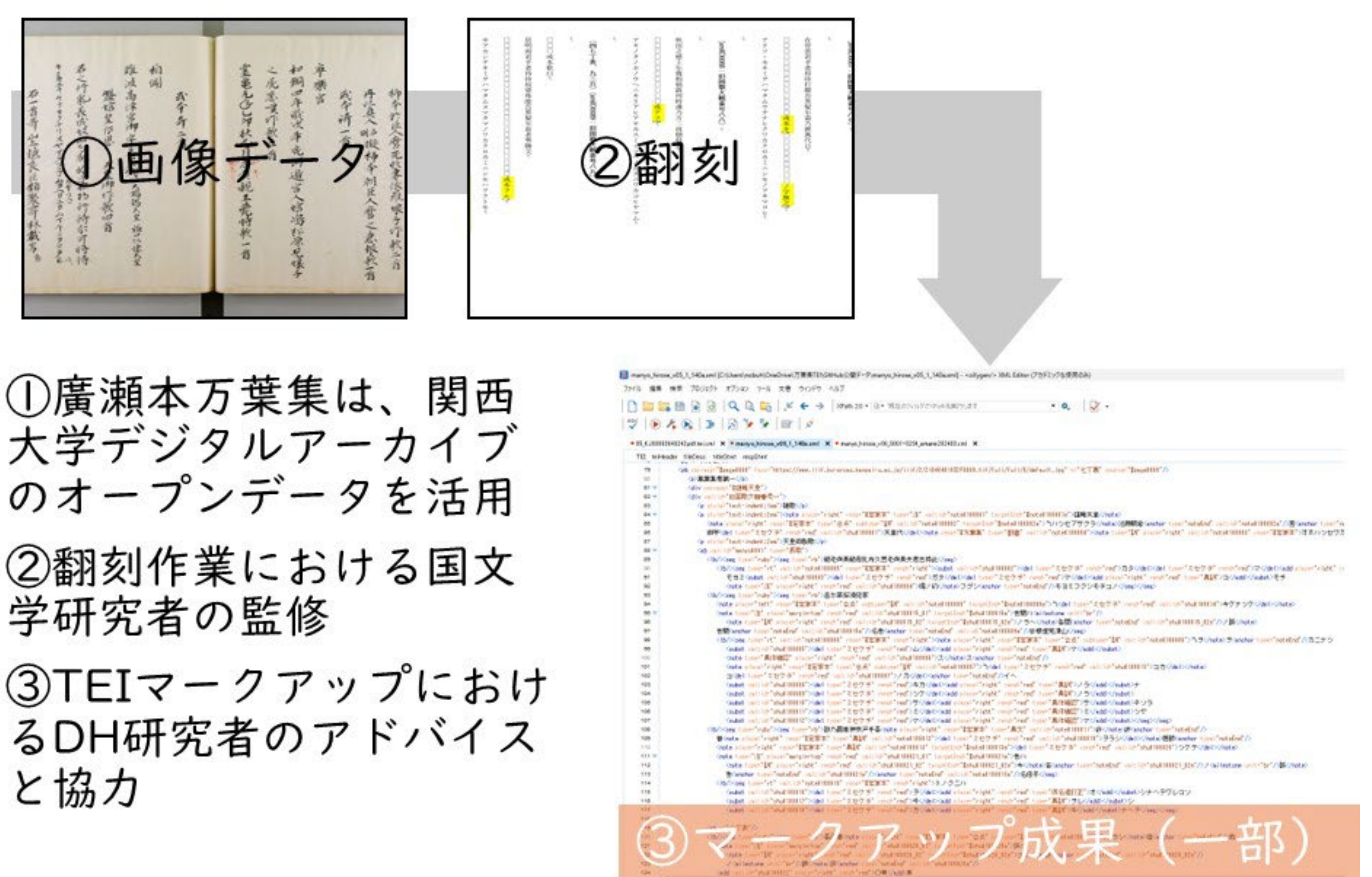
研究代表：菊池信彦（国文研）

国文学研究者：乾善彦（関西大）、茂野智大（筑波大）、Małgorzata (Gosia) K. Citko-DuPlantis (University of Tennessee, Knoxville)、青山英正（東京大）

人文情報学（DH）研究者：永崎研宣（慶應大学／人文情報学研究所）、吉賀夏子（阪大）

業務委託：合同会社AMANE

## 6. TEI/XMLマークアップの流れ



- ①廣瀬本万葉集は、関西大学デジタルアーカイブのオープンデータを活用
- ②翻刻作業における国文学研究者の監修
- ③TEIマークアップにおけるDH研究者のアドバイスと協力

## 7. その他の追加的成果

- 皆がXMLデータの解析に慣れているわけではないため
  - 分析手法の提供  
Google Colaboratoryを利用し、廣瀬本万葉集TEI/XMLデータから、情報を検索、抽出、（一部、可視化）まで行う教育用リソースを公開
  - 可視化手法を提供  
翻刻およびマークアップ結果を手軽に検索・閲覧できるビューワも開発（新バージョン開発中）
- AI等を活用を見込んだマークアップ作業のため、歌人人物データ作成も



分析手法  
(Google Colab)



ビューワ（イメージ）

## 8. 今後の展開

- 国文研の10年に及ぶ画像データベースプロジェクト成果「国書データベース」を踏まえ、これからの10年をかけて、データ駆動型人文学の創生に向けたデータ基盤を大学共同利用機関法人の一組織として構築していく
- TEI/XMLはそのデータ基盤の柱のひとつであり、国書データベースのメタデータ＋一部の古典籍テキストを対象に作成予定
- 「一部の古典籍」として廣瀬本万葉集を取り上げ、そのプロジェクトを紹介
- データ基盤の構築とともに、データ駆動型人文学の確立に向け、研究方法・環境を含めて普及を図っていく