様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １４ 日現在

機関番号：３４３１５
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１２５６７

研究課題名（和文）日本文化ディジタルアーカイブへの多言語統合アクセスの研究

研究課題名（英文）Research on multilingual integrated access to digital archives of Japanese culture

研究代表者
前田　亮（Maeda, Akira）

立命館大学・情報理工学部・教授

研究者番号：２０３５１３２２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では，研究代表者らが研究を進めてきた日本文化データベースの言語横断レコード同定技術を基盤とし，日本語で記述されたデータベースに対するバイリンガル検索技術，ディジタルアーカイブに対するグラフベースの情報推薦技術，蔵書印および落款印の文字認識および画像認識技術の各技術を統合することにより，日本のみならず世界に散在する日本文化ディジタルアーカイブに対する統合的な多言語情報アクセス環境を実現することを目指して研究を行った．その成果として，各技術について国際ジャーナルに論文が掲載され，また一部の技術については実稼働する日本文化ディジタルアーカイブに機能が搭載され，当初の目的を達成することができた．

研究成果の学術的意義や社会的意義
本研究の成果により，世界中に散在する人文系ディジタルアーカイブへの統合アクセスが可能となる．また，日本文化ディジタルアーカイブに対する海外の日本研究者からの容易な情報アクセスを支援する環境が実現する．これらより，これまでの人文系研究のように特定の分野や言語に閉じることなく，これらの壁を越えた，従来の人文学研究の方法論にとらわれない新たな研究手法への発展が期待でき，人文系研究の進展に貢献できると考えている．

研究成果の概要（英文）：This research is aimed at realizing an integrated multilingual information access environment for Japanese cultural digital archives that are hosted not only in Japan but also around the world. The research is based on the cross-language record linkage technology for Japanese cultural databases that has been developed by the principal investigator and collaborators, and it integrates the bilingual information retrieval technology for databases written in Japanese, the graph-based information recommendation technology for digital archives, and the character and image recognition of ownership and rakkan seals. As the result of the research, international journal papers have been published for each technology, and some of the technologies have been integrated into publicly open and working Japanese cultural digital archives, thus achieving our initial goal.

研究分野：人文情報学

キーワード：多言語処理　レコード同定　情報検索　情報推薦　文字認識　メタデータ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，国内外の図書館・博物館・美術館・文書館などにおいて，資料のディジタル化および公開が進んでいる．これらは，通常は各機関が個別にディジタルアーカイブとして公開を行っており，データベースによってユーザインタフェース，提供言語，メタデータスキーマなどが異なるのが通常であり，そのままでは統合利用は困難なのが現状である．一部では標準的なメタデータスキーマ（たとえば Dublin Core，CIDOC/CRM など）の適用も見られるが，現状では有効に活用されているとは言い難い．

複数ディジタルアーカイブの統合利用を可能としたシステムとして，人間文化研究機構の研究資源共有化システムや，国立国会図書館サーチ，Europeana などがこれまでに存在し，2019 年 2 月にはジャパンサーチ（試験版）が公開されるなど，統合利用の環境が整いつつある．しかし，これらは基本的に，事前に統合システム側でメタデータを収集する「ハーベスティング」もしくは横断検索のための登録作業が必要であり，多大な人的コストを要するのが現状である．

２．研究の目的

本研究では，研究代表者が 2016～2019 年度に行った科学研究費補助金 基盤研究（C）「多言語ディジタルアーカイブにおける言語横断レコード同定手法の研究」において新たに確立した，日本文化データベースの言語横断レコード同定技術を基盤とする．

また，研究代表者が研究分担者として 2015～2017 年度に参加した国文学研究資料館 研究開発系共同研究「新古典籍総合目録データベース」のマルチリンガル化対応のための基礎研究」において確立した，日本語で記述されたデータベースに対するバイリンガル検索技術を組み込むことにより，日本文化ディジタルアーカイブに対する海外の日本研究者からの容易な情報アクセスを支援する環境を構築する．

また，研究代表者が研究協力者と進めている，日本文化ディジタルアーカイブに対するグラフベースの情報推薦技術を用いることにより，日本文化の専門家のみならず初学者および一般の利用者に対しても，それぞれの目的に合った資料の発見を支援する環境を構築する．

さらに，研究代表者が研究協力者と進めている，蔵書印および落款印の文字認識および画像認識技術を応用することにより，日本文化資料にしばしば付される，古典籍などにおける収集機関の蔵書印および絵画や書などにおける作者の落款印や花押の情報を活用することで，これらの情報から得られる資料間に存在する潜在的な関係，さらには資料作者および資料収集者間の人的関係を明らかにし，近世・近代における日本の文化人ネットワークの分析を可能とする環境を構築する．

最終的には，これらの各技術を統合することにより，日本のみならず世界に散在する日本文化ディジタルアーカイブに対する統合的な多言語情報アクセス環境を実現することを目指した．

３．研究の方法

本研究では，これまで実現されていなかった日本文化ディジタルアーカイブに対する多面的かつ統合的な多言語情報アクセス環境を実現するため，主に以下の研究を行った．

（1）日本文化データベースの言語横断レコード同定技術

本研究は，世界の美術館，博物館，研究機関等で公開されている浮世絵データベースを対象として，メタデータが記述されている言語が異なっていても，同一の浮世絵作品のレコードを発見することを目的としている．同一言語で記述されたメタデータを対象として同一レコードを発見する技術は，レコード同定技術として古くから研究されているが，異なる言語を対象とした言語横断レコード同定は，研究代表者らが考案した独自の技術である．

言語横断レコード同定を実現するには，作品名などのメタデータ項目について，異なる言語間でのマッチングを行う必要があるが，海外の浮世絵データベースおける翻訳された作品名と，日本語による元の作品名のマッチングは容易ではない．

本研究では，まず，固有名詞の抽出および翻字を行うことにより，マッチングにおいて重要となる固有名詞の抽出精度を向上する手法を提案した．また，単語の意味のベクトル表現である単語分散表現を用いることで，メタデータ項目の意味的マッチングを行う手法を提案した．さらに，各言語の単語分散表現におけるベクトル空間のマッピングを学習することにより，翻訳手法に一切依存せずに異なる言語間でのメタデータの類似度を測る手法を提案した．

（2）日本文化ディジタルアーカイブに対するバイリンガル検索技術

本研究では，立命館大学アート・リサーチセンター（ARC）が提供する各種日本文化資源のデータベースに対して，日英の二言語による横断検索システムの実現を目指して研究を行った．

本システムの特徴の一つは，事前にメタデータを収集するのではなく，検索時にリアルタイムで各データベースにアクセスし，検索結果を動的に統合して表示する点にある．これにより，データベースの更新に対してタイムラグが発生せず，常に最新の情報を検索することが可能となった．また，データベース毎に異なるメタデータ項目を Dublin Core に基づく主要なメタデータにマッピングすることで，検索結果の統合表示を容易にした．

また，バイリンガル検索の機能として，専門用語辞書を用いた問合せ翻訳により，英語による問合せで日本語のメタデータを検索する言語横断検索の機能を実現した．さらに，日本語に不慣れな利用者による検索を容易にするために，日本語の形態素解析器および専門用語辞書により，日本語表記のメタデータからローマ字表記を自動生成して利用者に提示する機能を実現した．

本システムにより，複数データベースを一つの問合せで一度に検索することができ，さらに日本語でメタデータが記述されたデータベースに対して日英二言語での検索が可能となった．

（3）日本文化ディジタルアーカイブに対するグラフベースの情報推薦技術

既存の人文系データベースでは，検索機能はほぼすべてに実装されているが，情報推薦の機能を実装しているものは少ない．本研究では，多様な浮世絵データベースの利用者のニーズに応えるため，また，浮世絵データベースのさらなる有効活用を目指して，浮世絵データベースに対して情報推薦技術を適用する研究を行った．

本研究では，ARC 浮世絵ポータルデータベースのメタデータおよびアクセスログデータから，利用者と作品間の関係を示すグラフを構築し，これを基に推薦を行う手法を提案した．

構築したグラフから，リンク予測と呼ばれる手法を用いることで，利用者と作品間の潜在的な関係を推測することができる．これにより，ある利用者がある作品を好むかどうかを推測し，利用者が好むと思われる作品を推薦することが可能となった．

（4）蔵書印および落款の文字認識に基づく資料間・収集者間の関係分析技術

浮世絵作品には，しばしば絵師の落款印が捺されていることがある．これらには文字だけでなく図案が含まれる場合もあり，その解読は必ずしも容易ではない．本研究では，肉筆の浮世絵を主な対象として，落款印全体の検索および落款印中の各文字の検索の実現を目指して研究を行った．

落款印全体の検索では，落款印のデータベースとして「大日本書画名家大鑑 〈落款印譜編〉，荒木矩著，第一書房（1934）」を用い，ARC 浮世絵ポータルデータベースに含まれる浮世絵の落款印を入力として検索を行うシステムを実装した．落款印中の各文字の検索では，まず文字領域の自動抽出を行い，次に抽出された各文字について文字認識を行う手法を提案した．

４．研究成果

（1）日本文化データベースの言語横断レコード同定技術

言語横断型の単語分散表現を用いて機械翻訳に依存せずに異言語の浮世絵ディジタルアーカイブから同一作品を同定する手法について，言語横断型の単語分散表現に日本語特有の情報を用いることで精度を向上する手法を開発した．また，浮世絵を対象として，日英に加え新たに日本語とオランダ語間を対象とした実験を行い，日英言語間と同等の精度で言語横断レコード同定が可能であることを実証した．

さらに，言語横断型の単語分散表現に日本語特有の情報を用いることで分散表現のマッピング精度を向上する手法を提案し，Bilingual Lexicon Induction タスクにおいて特に日本語由来の単語について精度が向上することを実証した．本手法に関して，International Journal of Asian Language Processing に論文が掲載された．

（2）日本文化ディジタルアーカイブに対するバイリンガル検索技術

本研究で実現したバイリンガル検索システムは，立命館大学アート・リサーチセンター（ARC）の ARC ポータルデータベースの横断検索システム，ARC 所蔵資料公開データベースの横断検索システムとして，それぞれ一般に公開した．

また，日本文化ディジタルアーカイブに対する表現学習を用いた情報検索技術について研究を行い，色情報および構造情報を用いたクロスモーダル表現学習に基づく画像検索手法を提案した．英語による曖昧な言葉で表現された色のテキスト情報を用いることで，より人間の感覚に適応した新たな画像検索機能を実現した．本手法に関して，国際ジャーナル Future Internet に論文が掲載された．

（3）日本文化ディジタルアーカイブに対するグラフベースの情報推薦技術

浮世絵ディジタルアーカイブを対象とした，アクセスログからのリンク予測に基づくグラフベースの情報推薦システムを構築した．また，従来のアクセスログおよびメタデータに加え，画像特徴を用いた情報推薦システムを構築し，実稼働している浮世絵ディジタルアーカイブに実装した．ARC 浮世絵ポータルデータベースのアクセスログに基づくクエリ推薦手法の研究に関して，国際ジャーナル Applied Sciences に論文が掲載された．

（4）蔵書印および落款の文字認識に基づく資料間・収集者間の関係分析技術

浮世絵画像にしばしば捺される落款印および古典籍に捺される蔵書印に対して文字認識を行い，これらの情報から得られる資料間に存在する潜在的な関係や資料作者および資料収集者間の人的関係を明らかにし，近世・近代における日本の文化人ネットワークの分析を行う技術を開発した．これにより，落款印や蔵書印の情報を基に，浮世絵や古典籍の資料間に存在する潜在的な関係や資料作者および資料収集者間の人的関係を分析する新たな手法を提案した．本手法に関して，Journal of Data Mining and Digital Humanities に論文が掲載された．

また，印鑑画像を対象とした教師無し学習に基づく検索手法および文字セグメンテーション手法を提案し，印鑑特有の有効な特徴を明らかにした．また，単一事例表現学習に基づく古代文字認識手法を提案し，各文字に対して単一あるいは少数の字形データしか入手できない文字体系に対して，有効なデータ拡張手法および学習手法を提案した．本手法に関して，国際ジャーナル Data に論文が掲載された．

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 Jiayun Wang, Biligsaikhan Batjargal, Akira Maeda, Kyoji Kawagoe, Ryo Akama	４．巻 13
２．論文標題 Modified Conditional Restricted Boltzmann Machines for Query Recommendation in Digital Archives	５．発行年 2023年
３．雑誌名 Applied Sciences	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/app13042435	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Li Kangying、Wang Jiayun、Batjargal Biligsaikhan、Maeda Akira	４．巻 14
２．論文標題 Intuitively Searching for the Rare Colors from Digital Artwork Collections by Text Description: A Case Demonstration of Japanese Ukiyo-e Print Retrieval	５．発行年 2022年
３．雑誌名 Future Internet	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/fi14070212	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Li Kangying、Batjargal Biligsaikhan、Maeda Akira	４．巻 6
２．論文標題 A Prototypical Network-Based Approach for Low-Resource Font Typeface Feature Extraction and Utilization	５．発行年 2021年
３．雑誌名 Data	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/data6120134	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kangying Li, Biligsaikhan Batjargal, and Akira Maeda	４．巻 HistoInformatics
２．論文標題 Character Segmentation in Asian Collector's Seal Imprints: An Attempt to Retrieval Based on Ancient Character Typeface	５．発行年 2021年
３．雑誌名 Journal of Data Mining and Digital Humanities	６．最初と最後の頁 1-19
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yuting Song, Biligsaikhan Batjargal, and Akira Maeda	4．巻 30
2．論文標題 Learning Japanese-English Bilingual Word Embeddings by Using Language Specificity	5．発行年 2021年
3．雑誌名 International Journal of Asian Language Processing	6．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S2717554520500149	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

〔学会発表〕　計19件（うち招待講演　0件／うち国際学会　5件）

1．発表者名 Garmaabazar Khaltarkhuu, Biligsaikhan Batjargal, and Akira Maeda
2．発表標題 Text Classification of Modern Mongolian Legal Documents
3．学会等名 Sixteenth International Workshop on Juris-informatics (JURISIN 2022)（国際学会）
4．発表年 2022年

1．発表者名 Garmaabazar Khaltarkhuu, Biligsaikhan Batjargal, and Akira Maeda
2．発表標題 Text Classification of Modern Mongolian Documents Using BERT Models
3．学会等名 26th International Conference on Asian Language Processing (IALP 2022)（国際学会）
4．発表年 2022年

1．発表者名 苑 広媛，李 康穎，後藤 真，木村 文則，前田 亮
2．発表標題 『日本人名辞典』からの歴史人物情報の抽出： Few-shot学習による古文の固有表現抽出の試み
3．学会等名 人文科学とコンピュータシンポジウム
4．発表年 2022年

１．発表者名
三木 恵大，前田 亮，赤間 亮

２．発表標題
固有表現抽出手法を用いた古典文書からの歌舞伎役者情報の自動抽出

３．学会等名
第12回知識・芸術・文化情報学研究会

４．発表年
2023年

１．発表者名
苑 広媛，李 康穎，後藤 真，木村 文則，前田 亮

２．発表標題
事例ベース固有表現抽出を用いた古文からの歴史人物情報の抽出および活用

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）

４．発表年
2023年

１．発表者名
Ryoga Nagai and Akira Maeda

２．発表標題
Dataset Augmentation for Grammatical Error Correction Using Markov Chain

３．学会等名
World Congress on Engineering (WCE2021)（国際学会）

４．発表年
2021年

１．発表者名
Xintao Fang, Yuting Song, and Akira Maeda

２．発表標題
Joint Extraction of Clinical Entities and Relations Using Multi-head Selection Method

３．学会等名
2021 International Conference on Asian Language Processing (IALP 2021)（国際学会）

４．発表年
2021年

１．発表者名
東雲 陽美，青山 敦，前田 亮

２．発表標題
経営哲学に関するテキストにおける検索結果の多様性を考慮した検索システム

３．学会等名
人文科学とコンピュータシンポジウム

４．発表年
2021年

１．発表者名
川端 恵大，前田 亮，赤間 亮

２．発表標題
役者評判記を用いた役者情報の抽出

３．学会等名
人文科学とコンピュータシンポジウム

４．発表年
2021年

１．発表者名
Ryoga Nagai and Akira Maeda

２．発表標題
Sentence Pair Augmentation Approach for Grammatical Error Correction

３．学会等名
1st International Conference on Computational Intelligence for Engineering and Management Applications (CIEMA 2022)（国際学会）

４．発表年
2022年

１．発表者名
Jiayun Wang, Biligsaikhan Batjargal, Akira Maeda, Kyoji Kawagoe, and Ryo Akama

２．発表標題
Making Ukiyo-e Easier to Discover: A Recommender System for Digital Archives

３．学会等名
Digital Humanities 2020

４．発表年
2020年

１．発表者名
Yuting Song, Biligsaikhan Batjargal, and Akira Maeda

２．発表標題
Finding Identical Ukiyo-e Prints across Databases in Japanese, English and Dutch

３．学会等名
Digital Humanities 2020

４．発表年
2020年

１．発表者名
Kangying Li, Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama

２．発表標題
Toward Exploring Artist Information from Seal Images in Ukiyo-e Collections

３．学会等名
Digital Humanities 2020

４．発表年
2020年

１．発表者名
Yuting Song, Biligsaikhan Batjargal, and Akira Maeda

２．発表標題
A Preliminary Attempt to Evaluate Machine Translations of Ukiyo-e Metadata Records

３．学会等名
The 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)

４．発表年
2020年

１．発表者名
Kangying Li, Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama

２．発表標題
Artwork Information Embedding Framework for Multi-source Ukiyo-e Record Retrieval

３．学会等名
The 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)

４．発表年
2020年

１．発表者名
Li Kangying, Batjargal Biligsaikhan, 前田 亮, 赤間 亮

２．発表標題
浮世絵レコードのクロスモーダル多言語横断検索に向けて：Multilingual-BERTによる作品情報の特徴埋め込み抽出の試み

３．学会等名
第10回知識・芸術・文化情報学研究会

４．発表年
2021年

１．発表者名
王 嘉韻, Batjargal Biligsaikhan, 前田 亮, 川越 恭二, 赤間 亮

２．発表標題
深層学習モデルに基づく浮世絵画像検索システムの開発

３．学会等名
第10回知識・芸術・文化情報学研究会

４．発表年
2021年

１．発表者名
永井 涼雅, 前田 亮

２．発表標題
マルコフ連鎖モデルを用いた文章校正のためのデータ拡張

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム（DEIM2021）

４．発表年
2021年

１．発表者名
FANG Xintao, SONG Yuting, 前田 亮

２．発表標題
Joint Entity and Relation Extraction from Clinical Records Using Pre-trained Language Model

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム（DEIM2021）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	バトジャルガル　ビルゲサイハン (Batjargal Biligsaikhan) (30725396)	立命館大学・総合科学技術研究機構・助教 (34315)	
研究分担者	ＳＯＮＧ　Ｙｕｔｉｎｇ (Song Yuting) (50849388)	立命館大学・情報理工学部・助教 (34315)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関