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　３ 年　 ６ 月　 ８ 日現在

機関番号：８４６０４
研究種目：挑戦的研究（萌芽）
研究期間：2019～2020
課題番号：１９Ｋ２１６４３

研究課題名（和文）機械学習による画像自動分類を活用した考古学ビッグデータの構造化と情報探索への適用

研究課題名（英文）Automatic classification of images using machine learning to structure archaeological big data and enhance information retrieval

研究代表者
高田　祐一（Takata, Yuichi）

独立行政法人国立文化財機構奈良文化財研究所・企画調整部・研究員

研究者番号：５０７０８５７６

交付決定額（研究期間全体）：（直接経費）　　4,400,000円

研究成果の概要（和文）：本研究では、膨大な情報資産を「考古学ビッグデータ」と捉え、機械学習により構造化を進めることにより流通性と再利用性の向上をはかる。
2019年度は、報告書デジタルデータから遺物図面・遺物写真等の種類に大別する教師データを作成した。その教師データをもとに機械学習による画像自動抽出プログラムで、報告書デジタルデータから類似画像を大量に抽出するテストプログラムを実装した。2020年度は、プログラムと教師データを活用し、PDFから82万件の画像を自動抽出した。その画像群からさらに石器の種別ごとの教師データ54種類を作成した。機械学習にて類似度を算出し、石器種別ごとに類似画像を表示できるようになった。

研究成果の学術的意義や社会的意義
考古学は蓄積型の学問である。これまでの調査研究によって膨大な文字情報と画像情報が蓄積されているが、標準化・構造化が進まず、情報の体系的な検索と再利用性に課題がある。本研究では、膨大な情報資産を「考古学ビッグデータ」と捉え、機械学習により構造化を進めることにより流通性と再利用性の向上を図った。データ探索（データマイニング等）の基盤構築を目指した。主に報告書図面データを対象に機械学習にて類似度を算出し、石器種別ごとに類似画像を表示できるようになった。大量データから研究に有意な情報探索をできるようになった意義は大きい。

研究成果の概要（英文）：In this research, we regard the vast information assets as "archaeological big data", and aim to improve their distributability and reusability by promoting their structuring through machine learning. In FY2019, we created teacher data from the digital data of reports, which are roughly classified into types such as drawings of artifacts and photographs of artifacts. In FY2020, the program and the teacher data were used to automatically extract 820,000 images from the PDF. From this set of images, 54 types of teacher data for each type of stone tool were created. By calculating the degree of similarity using machine learning, it is now possible to display similar images for each type of stone tool.

研究分野：人文情報学

キーワード：データベース　考古学　画像認識　ビッグデータ　自動分類　機械学習　情報探索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

考古学は蓄積型学問であり、調査報告や知見など情報が増加していくことが学問の深化につながる。1970 年台・80 年台においては膨大な情報が毎年生成されるようになり、適切な情報管理が困難になりつつあった。そこで多量の情報に対処しうる情報処理システムの確立が叫ばれていた（岩本圭輔 1977,田中琢 1982,1988）。80 年台以降、さらなる埋蔵文化財保護行政の着実な推進により年間約 1400～1700 冊の発掘調査報告書が刊行され、これまでに約 20 万冊程度の報告書があると言われている。もはや印刷物の報告書を人間がすべて閲覧することは不可能な量となっている。

近年は報告書の電子公開事業が進み、その情報基盤である全国遺跡報告総覧には膨大なデジタルデータが登録され「考古学ビッグデータ」を形成している。全国遺跡報告総覧は、まだ全ての報告書が登録されていないものの日本全国の報告書ＰＤＦを登録している点において、画期的である。報告書発行機関約 564 機関がデータ登録し、約 28000 冊の報告書、文字数約 22 億文字、ページ数約 358 万ページを誇る（2021 年 6 月時点）。しかし、単純なテキスト検索のみでは、検索結果にノイズを多数含むため、「情報が多すぎて探せない」という情報爆発の弊害が起きている。せっかくの膨大な蓄積を十分に活かすことができていないのが現状である。その原因の一つに、人文系学問の共通課題である非構造化データ問題がある。膨大なデータが構造化されていないため、体系的な再利用が困難となっている。

２．研究の目的

本研究は、膨大な情報が蓄積され様々な可能性を持つものの非構造化データであるゆえに活用されていない考古学ビッグデータを対象に機械学習によって構造化し、データの再利用性を向上させることで、考古学研究の深化、社会への情報流通性を向上させることを目的とした。そのために、考古学ビッグデータの特性分析（目的Ⅰ）、考古学情報の根幹である画像データの自動分類と構造化（目的Ⅱ）、情報探索機能への実践適用（目的Ⅲ）の３点を目的とした。全国遺跡報告総覧に登録されている考古学ビッグデータを対象に報告書に閉じ込められているデータの構造化を図った。今回は画像データを対象とした。

３．研究の方法

上記目的を達成するために、以下３つの方法にて研究を推進した。

【方法１】考古学ビッグデータの特性分析（目的Ⅰ）

発掘調査では、大量の情報が様々な形式で記録され、調査成果として報告書にまとめられる。報告書には、テキスト、画像、数値データなどが混在している。これらの情報が生まれるプロセスや量などを定性的・定量的に分析する。方法３の情報探索への適用において、効果を高めるためにも必須の工程である。

【方法２】機械学習による画像自動分類（目的Ⅱ）

○準備：報告書のＰＤＦには様々な遺物や遺構の画像データ（写真・図面）が収録されている。ＰＤＦファイルは、印刷物のレイアウトを継承したまま電子化できるなどメリットが多いが、データ自体は構造化されていない。ＰＤＦファイルから画像データを種別ごとに自動抽出するには、機械学習による画像自動抽出と分類を行うための教師データが必要である。したがって、「考古学ビッグデータ」から、遺物図面・遺物写真・遺構図面・遺構写真の 4 種類に大別する教師データを作成する。

○自動抽出・自動分類：作成した教師データをもとに機械学習のソフトウェアライブラリを使用してＰＤＦから画像を抽出する。上記４種別にて抽出したのちに土器、埴輪、陶磁器など詳細の種別ごとに自動分類を行い、さらに形状が類似しているものを自動分類する。既に軒丸瓦を対象として自動抽出の実験は一部成功している。

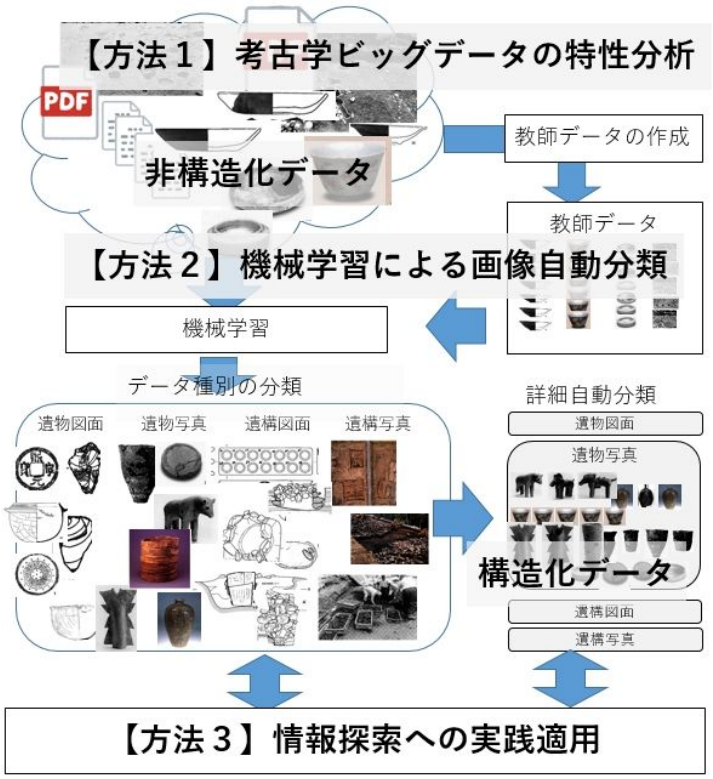

図１　研究計画フロー図

【方法３】情報探索への実践適用（目的Ⅲ）

考古学研究において、重要な作業は類例の調査である。類例との比較検証を積み上げることで新たな知見となる。画像を検索キーとする画像類似検索や、検索結果の絞り込みにメタデータを活用することで、必要な情報をより高度に選別できる情報探索法を検討し、実践適用する。

４．研究成果

報告書の電子公開は主にＰＤＦファイルによって実現されている。ＰＤＦファイルは、印刷物のレイアウトを継承したまま電子化できるなどメリットが多いが、データ自体は構造化されていないため、機械可読性は低い（高田 2020）。まずは 2020 年度に作成した教師データをもとに機械学習のソフトウェアライブラリを使用して、ＰＤＦファイルから画像 82 万件を自動抽出した。写真、図面が混在し、遺構および遺物すべての画像データであるため、まずは石器の図面画像のみを人間で選別した。

石器種別ごとの教師データ作成　石器の図面画像をさらに 54 に種別設定し、種別ごとに教師データを作成した（図２）。作業には野口淳・国武貞克・森先一貴の協力を得た。種別は次の通り。スタンプ形石器、ストーン・リタッチャー、ナイフ形石器、両面礫器、両面調整石器、円形掻器、削器、剥片、剥片石核、台形様石器、台形石器、台石、尖頭器、岩偶、彫器、御物石器、打製石斧、打製石斧（直刃斧）、抉入石器、掻器、敲石、斧形石器、有舌尖頭器、有茎石器、槌石、横長剥片、片面礫器、独鈷石、環石、石冠、石刃、石剣、石匙、石斧、石核、石棒、石皿、石篦、石銛、石錐、石錘、石鏃、砥石、磨り石、磨石、磨製石斧（両刃）、磨製石斧（片刃）、細石刃、細石刃核、細部加工剥片、縦長剥片、舟形石器、鋸歯状石器、錐

機械学習にて種別の類似度を算出　機械学習にて個別の石器画像ごとに 54 の種別の類似度を算出した（図３）。数値が高い種別ほど類似していることを示す。この工程によって石器種別ごとに分類できたことになる。

類似画像の表示　当該画像に類似している画像を表示させる機能も開発した。おおむね類似している画像を表示させることに成功した。しかし、一部関係のない画像も混入しており、精度の向上の余地は残る（図４・５）。

今後の展開　今後は、土器・瓦・木製品など遺物ごとに教師データを作成し、機械学習による分類と類似度算出を進める予定である。また、画像を検索キーとする画像検索も実装を検討する。遺物名が不明であっても画像で検索できるため、使途不明遺物への対応に有効となる。遺物名がわからない市民にとっても便利となる。

岩本圭輔 1977「埋蔵文化財関係用語の収集と整理」『奈良文化財研究所年報』

田中琢・佐原真 1993「切口上－エピローグ」『考古学の散歩道』

田中琢 1982「考古学、みかけだけのはなやかさ」『同朋』

高田祐一 2019「報告書のデータ量を推計する」『文化財の壺』(7)

高田祐一 2020「画像認識技術の文化財データへの適用実験」奈良文化財研究所紀要 2020

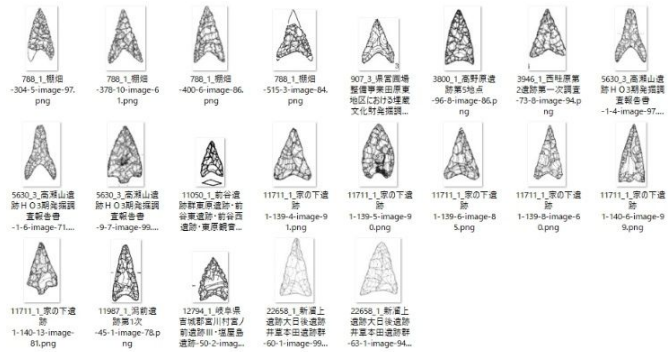

図２　石鏃の教師データとして選別した図面画像

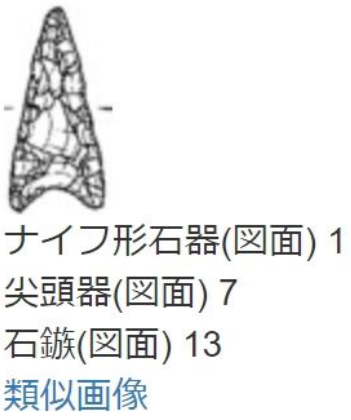

図３　機械学習にて各種別の類似度を算出

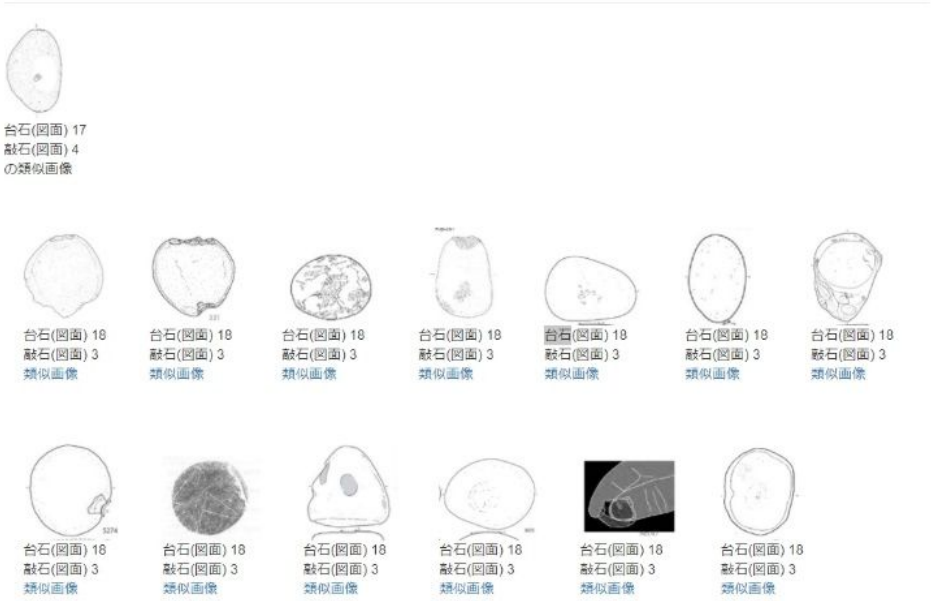

図４　台石の類似図面画像

図５　石匙の類似図面画像

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　0件／うち国際共著　1件／うちオープンアクセス　5件）

１．著者名 高田祐一 ， 金田明大 ， Dessislava Veltcheva	４．巻
２．論文標題 Prospects and potential for the comprehensive database of archaeological site reports in Japan	５．発行年 2019年
３．雑誌名 The ARIADNE Impact	６．最初と最後の頁 175-185
掲載論文のDOI（デジタルオブジェクト識別子） 10.5281/zenodo.3476712	査読の有無 無
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 該当する

１．著者名 高田祐一	４．巻 3
２．論文標題 考古学デジタルデータのアーカイブにおけるビジネスモデル－イギリスADS の事例から－	５．発行年 2021年
３．雑誌名 デジタル技術による文化財情報の記録と利活用3－著作権・文化財動画・GIS・三次元データ・電子公開－	６．最初と最後の頁 100-103
掲載論文のDOI（デジタルオブジェクト識別子） 10.24484/sitereports.90271	査読の有無 無
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 高田祐一	４．巻 3
２．論文標題 デジタル時代において文化財専門家に求められること	５．発行年 2021年
３．雑誌名 デジタル技術による文化財情報の記録と利活用3－著作権・文化財動画・GIS・三次元データ・電子公開－	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.24484/sitereports.90271	査読の有無 無
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 高田祐一,野口淳	４．巻 2020
２．論文標題 画像認識技術の文化財データへの適用実験	５．発行年 2020年
３．雑誌名 奈良文化財研究所紀要2020	６．最初と最後の頁 46-47
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 無
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 高田祐一,野口淳	４．巻 2021
２．論文標題 機械学習による石器図面画像の自動抽出と分類	５．発行年 2021年
３．雑誌名 奈良文化財研究所紀要2021	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） なし	査読の有無 無
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 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 高田祐一
２．発表標題 考古学・埋蔵文化財の情報プラットフォームとしての全国遺跡報告総覧－When Where Whatで検索できるシステムを目指して－
３．学会等名 日本考古学協会第87回総会研究発表
４．発表年 2021年

１．発表者名 武内樹治 ， 高田祐一
２．発表標題 文化財情報発信の現状と課題 －日本全国の文化財オープンデータ調査から考察する－
３．学会等名 日本情報考古学会第44回大会
４．発表年 2021年

１．発表者名 野口 淳 ， 高田祐一
２．発表標題 考古・埋蔵文化財空間データの可能性
３．学会等名 日本情報考古学会第44回大会
４．発表年 2021年

１．発表者名
Yuichi Takata ， Peter Yanase

２．発表標題
Opening the Vaults of Japanese Archaeology

３．学会等名
2021 AAS Annual Conference（国際学会）

４．発表年
2021年

〔図書〕　計2件

１．著者名 高田祐一（編著）	４．発行年 2020年
２．出版社 奈良文化財研究所	５．総ページ数 238
３．書名 デジタル技術による文化財情報の記録と利活用2 オープンサイエンス・データ長期保管・知的財産権・GIS	

１．著者名 高田祐一（編著）	４．発行年 2021年
２．出版社 奈良文化財研究所	５．総ページ数 160
３．書名 デジタル技術による文化財情報の記録と利活用3－著作権・文化財動画・GIS・三次元データ・電子公開－	

〔産業財産権〕

〔その他〕

全国遺跡報告総覧 https://sitereports.nabunken.go.jp/ja

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	野口　淳 (Noguchi Atsushi) (70308063)	独立行政法人国立文化財機構奈良文化財研究所・埋蔵文化財センター・客員研究員 (84604)	

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