画像認識技術の文化財データへの適用実験

はじめに　考古学は蓄積型の学問であり、発掘調査報告書（以下、報告書）は、重要な基礎資料である。しかし、報告書は戦前含め推計125,000冊あり膨大にあるため、過去の蓄積に適切にアクセスしにくいという課題もある[1)]。この課題を解決するためには、「発掘調査報告書」自体を分析し、掲載情報に適切にアクセスできる方法を考える必要がある。考古学においては、遺構・遺物そのものが研究対象であるため、画像情報（図面・写真）は重要である。本稿では、報告書に掲載された膨大な画像にアクセスするための画像認識技術の試行を報告する。

1980年代の実験例　80年台には画像を対象にした研究に挑戦している。田中琢が手掛けた「実験的研究は、画像処理に関するものである。出土遺物からその特徴を抽出し、それによって、型式の識別や遺物の対比をコンピュータでやろうという」野心的な取り組みであったが、「中絶」してしまった[2)]。しかし近年の画像認識技術の一般化によって、機械学習に使用できるソフトウェアライブラリがオープンソースで公開されるなど、技術環境が整ってきた。また報告書データベースである全国遺跡報告総覧（以下、遺跡総覧）の整備によって報告書電子データの蓄積が進んでいる。ソフトウェアとデータが揃ったことで、新たな挑戦が可能となった。

発掘調査報告書の画像数　これまでに発行された報告書に含まれている画像数は、どのくらいだろうか。画像数を推計するために、遺跡総覧に登録されている兵庫県教育委員会発行の兵庫県文化財調査報告シリーズ８から500までの347件を対象に画像数を集計した。報告書内の画像は写真（全景・風景・遺構・遺物など）、図面（拓本含む）・地図、表・グラフに分類し集計した。報告書347件の写真合計は18,859件、図面合計は22,682件、表・グラフの合計は2,271件となった。画像すべてで43,812件となった。１冊あたりでみるとそれぞれの中央値がページ数104、写真32、図面40、表・グラフ４となっており、これが兵庫県文化財調査報告シリーズの報告書で一般的な画像量といえる。これに報告書総数125,000件を乗算すると日本全体で、写真400万件、図面500万件、表・グラフ50万件となり、総合計が950万（9,500,000）件と推計できる。

考古学ビッグデータの構造化　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。人間可読性は高いが、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。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。それをそのままPDF化するため、様々な電子データが混在した非構造化データとなる。コンピュータがデータ処理するためにはデータを構造化する必要がある。その構造化を人間が手作業で実施するには膨大なコストがかかるため、実現可能性は低い。よって、膨大なPDFファイルからデータ属性ごとに自動で構造化する技術が求められる。PDFファイルから画像データのみを自動抽出し、種別ごとに自動分類するには、機械学習による画像自動抽出プログラムと分類するための教師データが必要である「考古学ビッグデータ」から、遺物図面・遺物写真・遺構図面・遺構写真等の種類に大別する教師データを作成した（図62）。

機械学習による画像自動分類　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から画像を抽出した。機械学習によって、軒丸瓦を対象とする自動抽出の実験は一部成功している（図63）。

図62　考古学ビッグデータ画像データの構造化モデル

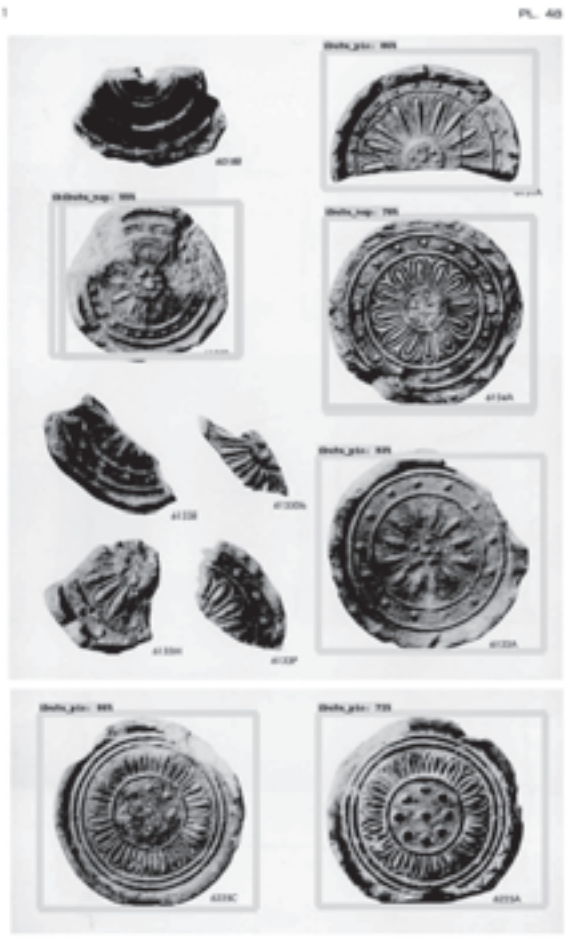

図63　軒丸瓦の自動抽出結果
（枠線部が軒丸瓦と識別できた）

図64　報告書群から自動抽出した遺物写真

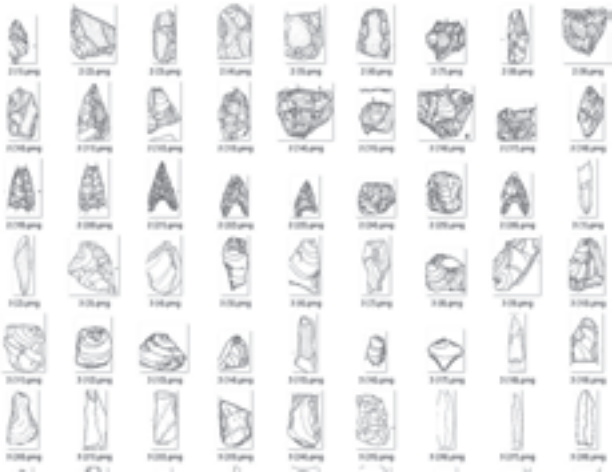

図65　報告書群から自動抽出した遺物実測図

完形はもちろんのこと、およそ50％強の残存率で軒丸瓦と識別できた。ほか、遺構写真や遺物写真などの自動抽出も成功している（図64・65）。今後、上記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予定である。

情報探索への実践適用　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、実践適用を検討していく予定である。

おわりに　報告書掲載画像推計950万件を一人の人間が確認することはほぼ不可能である。さらに画像が累積していく未来になっては加速度的に閲覧困難になるだろう。画像認識技術など大量データに打ち勝つための手段が必要である。さらに大量のデータならではの考古学研究が可能となれば、研究手法のブレイクスルーとなる可能性を秘めている。

なお、本稿は、研究課題19K21643の成果の一部である。

（高田祐一・野口　淳）

註

1）高田祐一「報告書のデータ量を推計する」『文化財の壺』7、2019。

2）田中琢「ある考古学研究者のパーソナルなコンピュータ史」『人文科学データベース研究』人文科学とデータベース研究刊行会、1988。