文化財ビッグデータと情報検索の可能性

はじめに　全国遺跡報告総覧（以下、遺跡総覧）には、遺跡情報を中心に膨大な文化財デジタルデータが集積されつつある。類似報告書の自動提示や普及活用イベント周知への応用などデータ集積によってこれまで実現できなかった新たな活用方法も生まれている。デジタル時代の新たな情報検索の可能性について報告する。

蓄積型の学問　考古学や歴史学はデータ蓄積型の学問であり、調査事例や研究成果の積み重ねによって深化していくという性格を持つ。そのため、調査研究に関するデータは蓄積されるほど有用だと考える。しかし、実際には情報爆発の弊害があり、情報が多すぎて管理できないというデメリットが生じている。このような傾向は、1970年代後半には、既に文化財関係資料が膨大となり「資料の全貌は、もはや誰にも把握しきれない。このため現在、研究、文化財・保護の仕事にたずさわる者が、過去の資料の蓄積を適切に選択して利用するのは、大変に難しいという状況にあり、将来この傾向がさらに甚だしくなることは目にみえている」と指摘されている[1)]。80年代には「多量の考古学資料の蓄積、厖大な情報量が、そのまま素晴らしい研究成果を生むものになっているとはいいがたい」、その解決のために「発掘調査のもたらす多量の情報に対処しうる情報処理システムの確立」が必要と指摘された[2)]。1994年には、報告書の内容を要約した抄録を報告書に添付する行政的取り組みが開始した。その後、抄録のデジタルデータを全国的に収集し、データベース化することも開始した。抄録には、遺跡の位置情報、時代情報、遺構遺物の情報が掲載されており、時空間情報で必要とする情報にピンポイントでアクセスできる画期的で重要な取り組みとなった。

全国遺跡報告総覧の展開　2008年度から2012年度にかけ、島根大学を中心とした全国の21の国立大学が連携して取り組んだ「全国遺跡資料リポジトリ・プロジェクト」（以下遺跡リポジトリプロジェクト）では約１万4,000冊の発掘調査報告書が電子化され、年間約50万件のダウンロードがあるなど、活発に利用され大きな成果をあげた。また、報告書全文に対し、１回でテキスト検索できるようになったことで、膨大な情報から網羅的に検索できる手段を確立した点において画期となった。しかし、各大学のサーバの老朽化など、プロジェクトの継続に課題があった。そこで、奈文研では、発掘調査報告書のメタデータを提供し、遺跡リポジトリプロジェクトで共同研究してきた経緯もあって、各大学21の遺跡リポジトリシステムと報告書の電子データを統合し、奈文研へ移管することとした。2015年６月から奈文研が全国遺跡報告総覧と改称し運用している。このような経緯もあって、大学・自治体・法人調査組織・学会等と共同推進する事業となっている。

統計的自然言語処理技術の活用　報告書の情報は、テキストと画像（図面と写真）にて構成される。遺跡総覧にはテキストが約17億文字登録されている（2019年３月時点）。これらの膨大なテキスト情報を情報資源化し可視化するには統計的自然言語処理技術が有効である。遺跡総覧では既に次の機能を実装している。

① 報告書ワードマップ（頻出用語俯瞰図）。７億文字のテキストについて、どういった考古学関係用語が頻出しているかを可視化したもの。

② 各都道府県版　報告書特徴語ワードマップ（図36）。当該都道府県内にて頻出する用語（よく使われる用語は重要）かつ他都道府県では出現頻度が低い用語（希

図36　各都道府県版　報告書特徴語ワードマップ

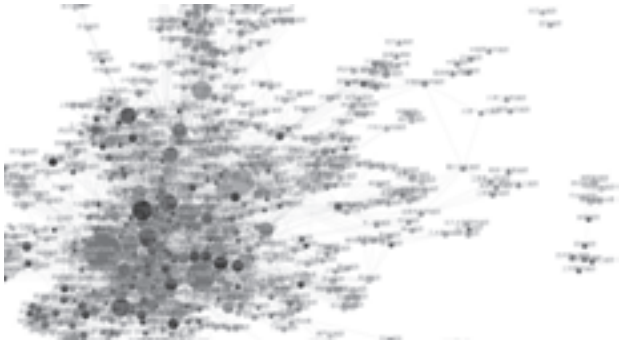

図37　遺跡（報告書）関係性ネットワーク図（旧石器遺跡）

少用語は重要）であることを勘案することで、当該都道府県の強い特徴を示す用語を可視化したもの。

③ 報告書毎の頻出用語と類似報告書とイベント情報の組み合わせ。頻出用語でおよその内容を把握し、類似の用語構成の報告書を自動提示することで、優先的に確認すべき報告書がわかる。そしてその報告書内容と類似しているイベントを表示することで、文化財事業と報告書の相乗効果を見込んでいる。

④ クロスリンガル機能。海外の研究者にとっては、日本考古学の成果に関心を示しながらも、言語の壁や報告書を手にとって閲覧できないという情報アクセスの問題がある。日本語を習熟しても、日本の考古学関係用語には多くの類語があり、それを全て覚えるのは困難であった。そこで文化財関係用語シソーラスを構築した。日英の考古学用語の対訳と日本語の考古学用語の類語をデータベース化し、英語自動変換機能を実装した。

他に、遺跡（報告書）関係性ネットワーク図（図37）を試行中である。報告書では、他遺跡への言及がある。それは当該遺跡を評価するために、周辺遺跡や類例などに言及するためである。この出現の組み合わせをカウントし、ネットワーク図として可視化した。2019年４月時点でまだ未公開である。

図38　軒丸瓦の画像自動抽出状況

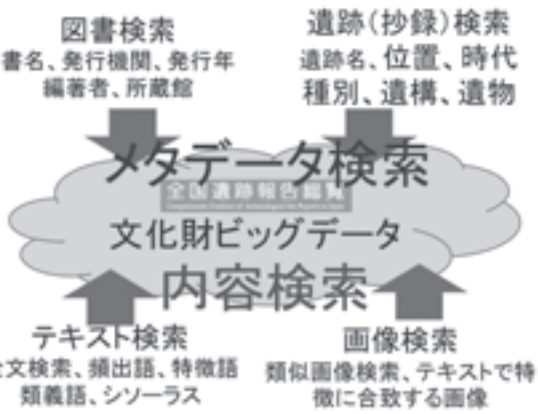

図39　文化財ビッグデータへの情報検索方法

人工知能による画像認識　　考古学においては、画像情報が重要である。80年代には画像を対象にした研究に挑戦している。「第二の実験的研究は、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」している[3)]。しかし近年の画像認識技術の向上によって、画像で画像を検索することも可能とする技術環境が整ってきた。機械学習技術を応用して実験適用中である（図38）。

おわりに　　文化財に関する情報を蓄積し、それにアクセスできる検索方法を確保する必要がある。これまでは図書としての検索や抄録検索であったが、テキストや画像そのものの内容を直接的に検索できる環境が構築できつつある（図39）。適切な検索手段の開発は、学術研究の促進や社会への情報発信に役立つと考える。

なお、本稿は、研究課題16H05881の成果の一部である。

（高田祐一）

註

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

2）田中琢「考古学、みかけだけのはなやかさ」『同朋』同朋舎出版、1982。

3）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。