考古学・埋蔵文化財にオープンでTidyなデータがなぜ必要なのか

野口淳（金沢大学古代文明・文化資源学研究所客員研究員）

On Why Archaeology and Buried Cultural Property Management Need Their Data to Be Open and Tidy

Noguchi Atsushi（ISACCR, Kanazawa University）

・オープンデータ／Open data・整然データ／Tidy data・集計表／Summary tables
・一覧台帳／Inventory・再利用性／Reproducibility

1. 概要

考古学・埋蔵文化財の調査研究では多種多様な記録が作成され利用される。それらは、あらゆる成果の基礎であり根幹である。しかし記録そのもの、その原データが再利用可能なかたちで公開されることは必ずしも一般的ではない。

国内において実施される発掘調査の大多数が「記録保存」であるだけでなく、そもそも発掘調査自体が地中に埋蔵されている遺跡・文化財の状態を不可逆的に変更する行為である。調査記録、そのベースとなる原データを公開することは、そうした変更に対する代償であり、発掘調査に直接関わることのなかった第三者に情報を共有することで透明性や検証可能性を担保することに他ならない。

本稿では考古学・埋蔵文化財におけるオープンデータの必要性とその要件について概観する。

2. なぜオープンデータが必要なのか？

考古学・埋蔵文化財の調査研究に関する記録、原データのオープン化がなぜ必要なのか。

発掘調査等による考古学・埋蔵文化財の調査研究成果は、多くの場合「報告書」の形式で公刊される。それは調査記録そのものではなく、調査記録を編集し、要約し、所定の書式に整えたものである。当然、原データがそのまま公刊されることはほとんどない。

調査記録に関する情報の公開共有が、原データではなく、編集・要約された報告書の形式で行なわれてきた理由のひとつに、可読性、理解の容易さがあるだろう。

たとえば多くの報告書では、出土資料の一覧台帳ではなく、さまざまな属性にもとづき集計した結果が、表や文章で示される。一覧台帳から、特定の属性にもとづき点数などの情報を取得するためには、必要なデータを抽出し数え上げなければならない。多用される属性情報を予め集計しておくことは、利用者の利便性を高めるので理に適っている。

しかし一覧台帳が公開されず集計表だけが提示される場合、再集計による検証はできない。通常であれば、報告書を信頼し、再集計・検証を必ず行なうことはない。とは言え、ミスが全く起こらないという保証もない。実際、クロス集計表において、個別項目の数値と行列の集計値が一致しない事例はある。そうした場合、一覧台帳も公刊されていれば、個別項目の数値が間違っているのか、集計値が間違っているのか、あるいは両者とも間違っているのかを検証できる。一覧台帳が公開されないということは、起こり得るミスをへのバックアップ、サポートが全くないということと同義である。

たとえば「記録保存」というスキームにおいて、または学術的な信頼性の担保において、この状況はどこまで許容されるのだろうか。

3. 本質データと派生物を区別する

もちろん集計表やグラフは、調査成果の全体像や要約を把握するのに優れている。逆に、収録項目やデータ件数が多い一覧台帳は、人間にとっての可読性が低く、内容を理解することが困難な場合も少なくない。

しかし集計表やグラフは調査記録そのものではない。調査記録にもとづく原データを一定の操作により要約し、または可視化した、派生物である。ほとんどの場合、そうした派生物は不可逆的な操作を経ており、原データを再現することはできない。したがって、当たり前のことではあるが、派生物は原データと等価ではない。

前節で指摘した集計表のミスをめぐる問題は、派生物ではない、本質データとしての原データが公開共有されることで解決できる。つまり必要なのは、本質データの公開、オープン化であり、これにより再現性、検証可能性が担保される。

ところで、成果の公開が印刷物によってのみ行なわれていた時には、膨大な分量にもなりかねない一覧台帳を交換することは、費用対効果の面から見て過大な要求であったとも言える。無理に小さな文字で印刷されたものもあるが、人間にとっての視認性も低下する上に、スキャンしての OCR（Optical character recognition：光学文字認識）にも適さない。これでは、コストをかけて印刷する意味がないだろう。

しかしデジタル化・情報化社会においては、一覧台帳は必ずしも印刷物として提供される必要はない。これまでにも磁気ディスク・光ディスクに収録した一覧台帳を印刷物と同時に頒布する事例があった（例：長野県埋蔵文化財センター編 2000）。インターネットの大容量・高速通信が一般化してからは、データファイルを直接ウェブサイトから提供する事例も登場している[1)]。

技術・手段の点で、いまや本質データをオープン化することへの障壁は無い。やるか、やらないかだけの問題である。

4. オープンデータの意義は再利用性

データがオープンであるということ、すなわちオープンデータとは、「自由に使えて再利用もでき、かつ誰でも再配布できるようなデータ」である[2)]。もう少し具体的に「機械判読に適したデータ形式で、二次利用が可能な利用ルールで公開されたデータ」であり「人手を多くかけずにデータの二次利用を可能とするもの」[3)] という定義も挙げておく。

自由に使え、再利用できることをより促進するためには、取得したデータを再加工等せずにそのまま使えることが重要になる。先に挙げた秋田市の事例はプロプライエタリなファイル形式ではあるが「機械判読に適した」ものであり、ダウンロードしてそのまま表計算ソフト等に読み込み、検索・抽出・並べ替え・集計などの操作が可能である。

一方、一覧台帳が公刊されていても、印刷物や PDF の場合は、再利用できるとは言え容易ではない。印刷物はスキャンし OCR を行なうことで、ようやく機械（コンピューター）で処理が可能になる。し

第４表　出土土器類観察表

No.	種別	器種	出土番号（注記）	器高（cm）	口縁径（cm）	底径（cm）
1	須恵器	坏蓋	石室 No.2	(1.8)	(10.4)	–
	色調（土色帖）	胎土		外面調整	内面調整	
	にぶい黄色（2.5Y 6／4）	砂粒等いっさいない緻密な胎土、焼成良好		ロクロナデ、所々深い線条痕残る、上部回転ヘラケズリ	ロクロナデ	
	残存度	備考				
	体部完存（つまみ部欠損）	つまみ部欠損、痕跡有、土器 No.2と同じ作り、７世紀後半～末				

No.	種別	器種	出土番号（注記）	器高（cm）	口縁径（cm）	底径（cm）
2	須恵器	坏	石室 No.19	(2.8)	(9.4)	不明
	色調（土色帖）	胎土		外面調整	内面調整	
	にぶい黄色（2.5Y 6／4）	No.1と同じ、砂粒等いっさいない緻密な胎土焼成良好		ロクロナデ	ロクロナデ	
	残存度	備考				
	口縁～体部下半 1／8	No.1と同じ作り				

No.	種別	器種	出土番号（注記）	器高（cm）	口縁径（cm）	底径（cm）
3	須恵器	坏	前庭部 No.13	口縁（2.5）、底部（1.3）	(16.8)	不明
	色調（土色帖）	胎土		外面調整	内面調整	
	灰白色（2.5Y 7／1）	緻密で不純物なし		ロクロナデ	ロクロナデ	
	残存度	備考				
	口縁～体部下半 1／8 底部 1／8	底部に歪る破片が４～８点あるが、外面が磨滅、口縁との接点なし 底部は口縁と同一個体であるが磨滅が著しい、底部はヘラ切り				

No.	種別	器種	出土番号（注記）	器高（cm）	口縁径（cm）	底径（cm）
4	須恵器	フラスコ形長頸瓶	石室 No.12ほか	21.2	(9.4)	–
	色調（土色帖）	胎土		外面調整	内面調整	
	明灰黄色（2.5Y 5／2） 灰黄褐色（2.5Y 7／2）	2 mm 大の白色・黒色・褐混入白色礫は 5 mm 大のものもある、比較的緻密な胎土・焼成良好		ロクロナデ、体部下半～底部にかけてケズリ調整、頸部を中心に全体的に自然釉かかる、頸部に一条の沈線	ナデと球状の器形を閉める輪積み痕残る	
	残存度	備考				
	口縁～底部 1／3	自然釉 球状の器に頸部を接合、東海西部産、頸部径（4.2）、体部径（15.0）				

図１　個票形式の「観察表」（安曇野市教育委員会 2022：第４表）

かし印刷物向けのフォーマットに組まれたものは、デジタルデータとしてのPDFを含め、そのまま表計算ソフト等に読み込むことはできない。データファイルとして提供されている場合と比較して、再利用の容易ではなく、操作性は格段に低い。

たとえば単純な「1行1件」の一覧表形式であれば、若干の操作で表計算ソフト等に読み込み可能なデータを作成することができるかもしれない。しかし、個票形式で組まれたものは、そうした二次的な加工も困難になる（図1）。これは「人手を多くかけずにデータの二次利用を可能とするもの」からはかけ離れている。個票形式の表示に慣れ親しんできたそうには「読みやすさ」を訴求するかもしれないが、再利用という観点からは継承する理由のないレガシー形式である。

5. Excelの乱れは心の乱れ

前節では一例として個票形式の書式を挙げたが、他にも「見た目」を重視する編集・組版の慣行は多数ある。データの再利用を阻害するそうした慣行の問題を端的に指摘するのが「Excelの乱れは心の乱れ」である。『東京都データ整備マニュアル』[4]の策定において唱えられた標語[5]は、総務省による『統計表における機械判読可能なデータの表記方法の統一ルール』[6]を参照し、以下の4点を遵守すべき事項として掲げている（図2）[7]。

・1つのセルには1つのデータ
・セル結合は使わない
・数値に単位や記号を混ぜない
・空白を入れない/改行しない

実際、これらは一覧表形式でデータを掲載している発掘調査報告書でもよく見かける。しかし確実に、コンピューターによる判読と再利用を阻害するものである。人間の読者にとってのぱっと見の「見やすさ」を優先して、コンピューターによるデータ処理の利便性を失うことは、果たして誰にとってのどのような利益になるのだろうか。

このほか全角/半角英数字の混在[8]、組文字等の

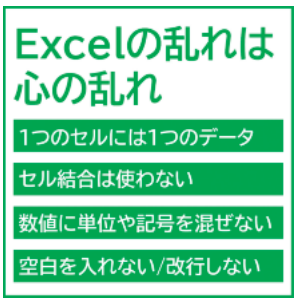

図2　Excelの乱れは心の乱れ

特殊記号や「機種依存文字」[9]の使用なども同根の問題がある。いまこれを改善せずに今後も継続することは、デジタル化・情報化が進む考古学・埋蔵文化財分野にとって「百害あって一利なし」である。

6. Tidyで再利用可能なデータを目指して

では、なぜ再利用性にこだわるのか。せっかく公開され利用可能な状態にあるデータを、実際に利用するにあたって利用者が毎回、手間をかけて利用可能な状態に加工することを繰り返すのは非効率だからである。そこにかける労力・コストを、他の必要な仕事に振り向ければ、より多くの新しい成果が期待できるだろう。逆にそれをしないことは、サンクコストを累積させ続けることに他ならない。

そして再利用性を低下させるのは、見た目を重視した編集・表組みだけにとどまらない。問題を回避するためには「整然データ（Tidy data）」の理解が必須である（西原2017）。

オープンデータと、その背後にあるオープンサイエンスの概念は、「データの所有権ではなく管理義務、分析過程の秘匿性より公開性、そして一般の人々の排除より包摂」を唱えるという新しい規範にもとづくものである（Marwick2019）。つまりオープンデータの推進は、再利用の利便性という実利的な側面だけでなく、学術分野における調査研究の透明性や公平性にもつながるものである。

【註】

1）秋田市「秋田市埋蔵文化財調査報告書『地蔵田遺跡　旧石器時代編』について」 https://www.city.akita.lg.jp/

kurashi/rekishi-bunka/1011795/1010787/1002234.html（2022/10/26閲覧）
秋田市「秋田市埋蔵文化財調査報告書『下堤G遺跡　旧石器時代編』について」 https://www.city.akita.lg.jp/kurashi/rekishi-bunka/1011795/1010787/1002236.html（2022/10/26閲覧）
いずれもMicrosoft Excelファイル（.xls形式）で出土石器全点の属性表が提供されている。

2）Open Knowledge Foundation オープンデータハンドブック日本語版「オープンデータとは何か?」http://opendatahandbook.org/guide/ja/what-is-open-data/（2022/10/26閲覧）

3）福岡市自治体オープンデータ「オープンデータとは?」https://www.open-governmentdata.org/about/（2022/10/26閲覧）

4）東京都デジタルサービス局「東京都データプラットフォーム　データ整備事業」https://www.digitalservice.metro.tokyo.lg.jp/society5.0/data_maintenance.html（2022/10/26閲覧）

5）https://twitter.com/miyasaka/status/1509595861956898817?s=20&t=wUMtuDgduyg_FGLO0J6djQ（2022/10/26閲覧）

6）総務省「統計表における機械判読可能なデータの表記方法の統一ルールの策定」 https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000186.html（2022/10/26閲覧）

7）NaoyaShimizzz「Excelの乱れは心の乱れ」https://github.com/NaoyaShimizzz/excel_disorder（2022/10/26閲覧）

8）文字幅を揃えるために1桁の数値は全角、2桁の数値は半角とするような編集・組版処理が行なわれる場合があるが、コンピューターによるデータ処理では全角と半角の英数字は別のものとして扱われるので絶対に避けるべきである。

9）環境依存文字とも。OSやソフトなどにより表示が一致しない、文字コードに互換性のないものを指す。今日では多くの文字・記号がUnicodeに収録されているが、完全な互換性が担保されていない領域については使用しないことが推奨される。たとえばローマ数字はUnicodeにも対応しているが、ラテン文字の使用が推奨される。

【引用文献】

Marwick, B.／高田祐一・野口 淳・P. Yanase 訳　2019「考古学における研究成果公開の動向 －データ管理・方法の透明性・再現性－」『デジタル技術による文化財情報の記録と利活用2』奈良文化財研究所研究報告24: 1-13 http://doi.org/10.24484/sitereports.69974

安曇野市教育委員会　2022『穂高古墳群C2号墳』安曇野市の埋蔵文化財26　http://doi.org/10.24484/sitereports.129425

長野県埋蔵文化財センター　2000『上信越自動車道埋蔵文化財発掘調査報告書16 信濃町内その2 信濃町データ編［CD-ROM］』長野県埋蔵文化財センター発掘調査報告書49　http://doi.org/10.24484/sitereports.8534

西原史暁　2017「整然データとは何か」『情報の科学と技術』67: 448-453　https://doi.org/10.18919/jkg.67.9_448