Archeologia e Calcolatori
35.2, 2024, 381-386
doi 10.19282/ac.35.2.2024.40

LEGACY DATA OR JUST ARCHAEOLOGICAL DATA?

1. Introduction

The proliferation of digital information resources is generating a phenomenon in every field of human knowledge commonly identified as 'datafication' (Mayer-Schonberger, Cukier 2013). Originally, the concept described the tendency to transform every aspect of life into data. Later, the term was used to define behaviours that pushed towards greater automation in handling large collections of data. 'Data is the new oil' has become over time the paradigm of information players (Google, Microsoft, Amazon, etc.) that produces added value on data collection and exploitation. The introduction of Big Data and, more generally, the need to train models for artificial intelligence, have modified the initial idea of 'datafication', which today coincide with a particular method of seeing, representing, discovering, and exploring information. Thus, data is replacing the methodologies aimed at producing, managing, and interpreting them.

The diffusion and use of data in different fields, such as scientific and educational, have led to a paradigm shift characterized by intensive data use and the possibility of automatically extracting and inferring knowledge from a vast collection of digital documents (Hey, Tansley, Tolle 2009). This new form of science, based on the fundamental activity of capturing, curating, and analysing data, aims not to propose simulative models but to enable the scientific exploration of nearly infinite collections of records available online, enhancing the analytical potential of each researcher. To access the materials needed for their research, researchers should sift through a digital library characterized by volumes, articles, reports, tables, images, drawings, photographs, sorted according to different standards, different languages, etc., and especially without a careful librarian capable of preserving memory and trace of the contents. Computational sciences increasingly support online research through the development of hardware and software systems aimed at integrating resources on the web, overcoming data fragmentation and heterogeneity; despite the encouraging technological innovations, many of the proposed solutions are partial and not entirely conclusive.

A significant impetus to the automatic processing of large amounts of data comes primarily from the implementation of new 'intelligence' techniques aimed at countering international terrorism through a broader sharing of decentralized databases (9/11 Commission Report, 2004). The need to increase the quality of data traffic control on the Internet, often consisting of unstructured information, has accelerated the development of algorithms capable of constructing aggregations of suspicious data (Biltgen, Ryan 2015).

The challenge in collecting and analysing data is assuming a central role in the relationship between emerging technologies, raw information available online, and the automatic extraction of content that can be grouped together. Innovation, therefore, looks towards a new interdisciplinary interaction aimed at overcoming the thin boundary between physical and virtual reality. Archaeological research has also been impacted by an explosion of digital data, digitised or digitalized data from new and recent discovery or from old archives. More generally, information technology has caused such a change in researchers' attitudes that it is thought that all archaeologists have become digital, although differences persist in skills and in access to and use of data (Morgan, Eve 2012). This contribution aims to explore a particular aspect in the process of re-elaborating digital data in archaeological research, namely that of legacy spatial data, which more generally we can identify in previous studies, digital or paper, structured or unstructured, often built with now outdated methodological approaches.

2. Data vs datafication in archaeology

Archaeological data plays a central role in research to the extent that we speak of the formation of the archaeological record to indicate which processes have produced that trace that the archaeologist subsequently records. While the theoretical and methodological debate seems to oscillate between very divergent opinions on the nature of excavation and its related archaeological documentation, some fixed points can be noted. Archaeology is characterized by poorly defined variables, often mistakenly thought of as data, derived from populations not always fully understood and from uncertain articulations between the entities whose logical relationships we seek to understand (Chippindale 2000). Archaeological research often moves in a marshy terrain characterized by uncertain boundaries, stagnant waters, muddy soil, and with a particular vegetation and fauna. For these reasons that M.B. Schiffer (1987) defines the archaeological record as the distorted reflection of an object that was once part of a more comprehensive behavioural system that we only partially reconstruct. According to I. Huvila (2017), data can be recorded and organized from a dozen different perspectives, and to emphasize the ambiguous and misleading nature of the variable in archaeological documentation, he coined the term MEAN (Miscellaneous Exceptional Arbitrary Nonconformist).

Examining the consistency of the digital archaeological record, J. Huggett (2022) has recently listed some of the main incongruities that can be synthesized as follows:

A – Tables, databases, relationships, texts can be incorporated within a structured data model, reducing variability, or eliminating those descriptive elements that do not fit the schema;
B – Raw data can be set aside in favour of processed data sets. The distinctions between primary, secondary, and tertiary data are lost by eliminating the distinction from what was originally collected to what was subsequently processed and interpreted;
C – The same data entities can be evidence of multiple phenomena.

Since data depends on the context of use and, at the same time, on users' beliefs, the results will always be highly complex, unstable, and unpredictable. All the cautions expressed by researchers about digital data and their use should warn archaeologists against uncritically accepting algorithms capable of automatically processing and aggregating large amounts of data. Only careful analysis of the entire digitization process can truly innovate the approach to excavation (Roosevelt *et al.* 2015), contributing, together with the increase in tools and sensors for data acquisition, to a paradigm shift (Huggett 2015; Schmidt, Maverick 2020). In the future, digitalization will simplify the work of archaeologists, who will be able to focus on examining more general and theoretical issues rather than organizing data. Despite the optimistic forecasts, outside of this promising scenario will remain the past excavations that preserve paper data or collections of digital records coded with old programs and according to outdated methods.

3. Legacy spatial archaeological data

Huggett (2018) has pointed out how the reuse of digital archives involves data aggregation that creates new values, which at the end of the process, however, can be more ambiguous and less transparent (Clarke 2016) in the absence of precise contextual data. The choice to use metadata and paradata associated with archives certainly contributes to increasing data understanding; however, the inherently unstable nature of any data integration process makes the Big Data scenario an objective with uncertain outcomes. Processing large amounts of archaeological records from different excavations and research necessarily entails a reconsideration of data recording methods and archive creation. Such caution necessarily increases when spatial data acquired at different times and with different methodologies are reused.

A trend towards the adoption of automated computer means to make efficient and effective use of large data sets was already present at the end of 1980s (Kvamme 1989). The pioneering use of databases to store and process large amounts of data was replaced by GIS systems, while in the same years CAD was established for the realization of topographic drawings and graphic documentation of excavations, allowing for rapid updating of plans and stratigraphy (Alperson-Afil 2019). The parallel development of GIS and CAD technologies, both based on a common numerical representation of information, has pushed industries towards greater integration between the two systems. However, despite the development of interoperable formats, migrating CAD drawings to GIS requires specific conceptualization, making automatic conversion impossible. Adapting CAD files for GIS applications, therefore, remains a challenging task, impossible to complete without substantial human intervention (Bibby, Ducke 2017). The difference between the two systems lies in the connection between geometric entities and the alphanumeric information associated with them. In particular, CAD produces complex graphical maps but ignores the non-spatial attributes associated with graphic entities. Therefore, reusing CAD data in GIS requires a reorganization of geometric and spatial information based on the identification and construction of objects that transform geometric primitives into semantic categories (walls, rooms, streets, buildings, etc.).

A test, conducted for the implementation of a GIS for the archaeological site of al-Balīd, ancient Zafar, in the Sultanate of Oman, confirmed the difficulty in designing and creating an automatic path for importing and managing a previous cartographic archive consisting mainly of CAD maps. Over the last 70 years, several teams have investigated the area of Zafar, an important Islamic-period maritime stopover located along the routes crossing the Indian Ocean (D'andrea 2021). In 1995, a German mission, led by M. Jansen (2015), was tasked with creating the archaeological park of the site. At the end of the work, a substantial digital archive was produced, including dozens of CAD files reproducing the archaeological area. In 2019, a research group from the University of Naples L'Orientale resumed investigations in the area with the aim, among others, of systematizing previous research and creating a GIS that would collect previous site plans and drawings. The review of the CAD archive started from the analysis of the survey of the Great Mosque, with the aim of verifying the accuracy of spatial information and the possible migration to GIS. The table associated with the spatial information present in QGIS lists 245 vectors that are not convertible automatically into architectural and structural elements of the religious building without human intervention. Only by reading the excavation reports is it possible to correctly identify the individual components of the mosque (walls, columns, stairs, thresholds, etc.) and build the corresponding objects in the GIS, referring, in some parts, to distinct phases of the structure's life cycle. The design of the GIS inevitably pushed towards a re-reading of the entire cartographic archive and, above all, of all the available archaeological documentation (graphic, photographic, and textual).

4. Conclusion

The experience gained in the project of converting the graphic archive of al-Balid, however partial, confirms the impossibility of designing an automatic treatment for the migration of CAD files into a GIS environment. The Archaeology Data Service has initiated a large-scale migration process of CAD files without, however, foreseeing a specific method for reusing digital drawings in different software and applications (Green *et al.* 2016).

The challenge for a correct conversion of plans, sections, and elevations must start from the identification of represented objects and not simple graphical entities. The category of spatial legacy data must be treated in the same way as a traditional cartographic archive, requiring a reading, also methodological, of all existing documentation. Only by following this path is it possible to correctly interpret the information drawn by the archaeologist, regardless of the format used, paper or CAD.

In the future, artificial intelligence will certainly provide new tools to automatically associate digital or handwritten texts with plans, making all this documentation readable by a machine (Fletcher 2023). However, pending the transformation of legacy data into 'reborn digital data', we still need to rely on the traditional methodology of spatial information processing, which places human experience at the centre of analysis.

Andrea D'Andrea
Dipartimento Asia, Africa e Mediterraneo
Università degli Studi di Napoli L'Orientale
dandrea@unior.it

REFERENCES

Alperson-Afil N. 2019, *Digitising the undigitized: Converting traditional archaeological records into computerized, three-dimensional site reconstruction*, «Journal of Graphic Information System», 11, 747-765 (https://www.scirp.org/journal/paperinformation?paperid=97426).

Bibby D., Ducke B. 2017, *Free and open-source software development in archaeology. Two interrelated case studies: gvSIG CE and Survey2GIS*, «Internet Archaeology», 43 (https://doi.org/10.11141/ia.43.3).

Biltgen P., Ryan S. 2015, *Activity-Based Intelligence: Principles and Applications*, Boston-London, Artech House.

Chippindale C. 2000, *Capta and data: On the true nature of archaeological information*, «American Antiquity», 65, 4, 605-612.

Clarke R. 2016, *Big data, big risks,* «Information Systems Journal», 26, 77-90 (https://doi.org/10.1111/isj.12088).

D'Andrea A. 2021, *Reconsidering the topography of al-Balid: A preliminary review of the graphical documentation*, «Annali Sezione Orientale», 81, 39-50 (https://doi.org/10.1163/24685631-12340110).

Fletcher E.C. 2023, *Creating a software methodology to analyze and preserve archaeological legacy data*, «Advances in Archaeological Practice», 11, 2, 139-151 (https//doi.org/10.1017/aap.2022.44).

Green K., Niven K., Field G. 2016, *Migrating 2 and 3D datasets: Preserving AutoCAD at the Archaeology Data Service*, «ISPRS International Journal of Geo-Information», 5, 44 (https://doi.org/10.3390/ijgi5040044).

Hey T., Tansley S., Tolle K. (eds.) 2009, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Washington, Microsoft Research.

Huggett J. 2015, *Challenging digital archaeology,* «Open Archaeology», 1, 1, 79-85 (https://doi.org/10.1515/opar-2015-0003).

Huggett J. 2018, *Reuse remix recycle: Repurposing archaeological digital data*, «Advances in Archaeological Practice», 6, 2, 93-104 (https://doi.org/10.1017/aap.2018.1).

Huggett J. 2022, *Data legacies, epistemic anxieties, and digital imaginaries in archaeology*, «Digital», 2, 2, 267-295 (https://doi.org/10.3390/digital2020016).

Huvila I. 2017, *Being FAIR. When archaeological information is MEAN: Miscellaneous, exceptional, arbitrary, nonconformist*, Presentation at the Centre for Digital Heritage Conference (Leiden 2017) (http://www.istohuvila.se/node/526).

Kvamme K. 1989, *Geographic Information Systems in regional archaeological research and data management*, in M.B. Schiffer (ed.), *Archaeological Method and Theory*, 1, Tucson, University of Arizona Press, 139-203.

Jansen M. 2015, *The archaeological park of Al-Baleed, Sultanate of Oman. Site atlas along with selected Technical Reports 1995-2001*, Muscat, Office of the Adviser to His Majesty the Sultan for Cultural Affairs.

Mayer-Schonberger V., Cukier K. 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, New York, Houghton Mifflin Harcourt.

Morgan C., Eve S. 2012, *DIY and Digital Archaeology: What are you doing to participate?*, «World Archaeology», 44, 4, 521-537.

Roosevelt C. H., Cobb P., Moss E., Olson B.R., Ünlüsoy S. 2015, *Excavation is ~~destruction~~ digitization: Advances in archaeological practice*, «Journal of Field Archaeology», 40, 3, 325-346 (https://doi.org/10.1179/2042458215Y.0000000004).

Schiffer M.B. 1987, *Formation Processes of the Archaeological Record*, Albuquerque, University of New Mexico.

Schmidt S.C., Marwick B. 2020, *Tool-driven revolutions in archaeological science*, «Journal of Computer Applications in Archaeology», 3, 1, 8-32 (https://doi.org/10.5334/jcaa.29).

The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks upon the United States, 2004, Washington, Government Printing Office (https://www.govinfo.gov/content/pkg/GPO-911REPORT/pdf/GPO-911REPORT.pdf).

ABSTRACT

The world of research is currently undergoing a profound transformation, characterized by the extensive use of digital data available online. To optimize the utilization of these resources, artificial intelligence offers researchers several tools capable of aggregating both structured and unstructured information. The need to train algorithms to enhance the use of artificial intelligence techniques in data classification has led to the creation of structured datasets. However, it is not always possible to fully automate the transfer of data to more modern environments without substantial human intervention, aimed at extracting the implicit knowledge present in digital data. The category of CAD data appears to be particularly challenging in terms of automated management of spatial resources. The use of graphical entities for digital drawings, without semantically identified components, makes automatic conversion into GIS extremely complex. The paper is based on a partial test conducted on a cartographic archive that has been formed over 70 years of field research, aiming to demonstrate the importance of prioritizing legacy spatial data, both digital and non-digital, as archaeological data.