DOI: 10.26314/GARR-Conf17-proceedings-03

Big Archaeological Data. The ArchAIDE project approach

Francesca Anichini, Gabriele Gattiglia

Università degli Studi di Pisa

Abstract. Digitisation has changed archaeology deeply and has increased exponentially the amount of data that could be processed, but it does not by itself involve datafication, which is the act of transforming something (objects, processes, etc.) into a quantified format, so they can be tabulated and analysed. Datafication fits a Big Data approach and promises to go significantly beyond digitisation. To datafy archaeology would mean to produce a flow of data starting from the data produced by the archaeological practice, for instance, locations, interactions and relations between finds and sites. The ArchAIDE project goes exactly in this direction. ArchAIDE is a H2020 funded project (2016-2019) that will realise a tool for recognising archaeological potsherds; a web-based real-time data visualization to generate new understanding; an open archive to allow the archival and re-use of ar-chaeological data. This process would move archaeology towards data-driven research and Big Data.

Keywords. Archaeology, Digitisation, Datafication, data-driven research, Big Data

Introduction

Data are what economists call a non-rivalrous good, in other words, they can be processed again and again and their value does not diminish (Samuelson, 1954). On the contrary, their value arises from what they reveal in aggregate. On the one hand, the constant enhancement of digital applications for producing, storing and manipulating data has brought the focus onto data-driven and data-led science even in the Humanities, on the other hand, in recent decades, archaeology has embraced digitisation. In recent years, archaeologists began to ask to themselves if a Big Data approach can be applied to archaeology from both a theoretical and practical point of view (Gattiglia 2015).

For a better understanding of the general concept of Big Data, we adopt the definition proposed by (Boyd et al. 2012): "Big Data is less about data that is big than it is about a capacity to search, aggregate, and crossreference large data sets". In other words, Big Data's high volume, high velocity, and high variety do not have to be considered in an absolute manner, but in a relative way. As suggested by (Mayer-Schönberger et al. 2013), using Big Data means working with the full (or close to the full) set of data, namely with all the data available from different disciplines that can be useful to solve a question (Big Data as All Data). This kind of approach permits to gain more choices for exploring data from diverse angles or for looking closer at certain features of them, and to comprehend aspects that we cannot understand using smaller amounts of data.

1. Datafication

Digitisation has changed archaeology deeply increasing exponentially the amount of data that could be processed, but from a more general point of view the act of digitisation, i.e. turning analogue information into computer readable format, does not by itself involve datafication. Datafication promises to go significantly beyond digitisation, and to have an even more profound impact on archaeology, challenging the foundations of our established methods of measurement and providing new opportunities. To datafy means to transform objects, processes, etc. in a quantified format so they can be tabulated and analysed (Mayer-Schönberger et al. 2013). Moreover, a key differentiating aspect between digitisation and datafication is the one related to data analytics: digitisation uses data analytics based on traditional sampling mechanisms, while datafication fits a Big Data approach and relies on the new forms of quantification and associated data mining techniques, that permit more sophisticated mathematical analyses to identify non-linear relationships among data, allowing us to use the information, for instance, for massive predictive analyses. In other words, to datafy archaeology would mean to produce a flow of data starting from the data produced by the archaeological practice, for instance, locations, interactions and relations between finds and sites. A flow of data that the archaeological community should have available.

2. ArchAIDE project

The ArchAIDE project goes exactly in this direction. ArchAIDE is a three-year (2016-2019) RIA project, approved by EC under call H2020-REFLECTIVE-6-2015. The project consortium is coordinated by the University of Pisa with the MAPPA Lab, and includes a solid set of Human Sciences partners (University of Barcelona, University of Cologne and University of York), some key players in ICT design and development (CNR-ISTI and Tel Aviv University), two archaeological companies (BARAKA and ELEMENTS) and one ICT company.

The work of the project includes the design, development and assessment of a new software platform offering applications, tools and services for digital archaeology. This framework, that will be available through both a mobile application and a desktop version, will be able to support archaeologists in recognising and classifying pottery sherds during excavation and post-excavation analysis. The system will be designed to provide very easy-to-use interfaces (e.g. touch-based definition of the potsherd profile from a photograph acquired with the mobile device) and will support efficient and powerful algorithms for characterisation, search and retrieval of the possible visual/geometrical correspondences over a complex database built from the data provided by classical 2D printed repositories and images. We thus plan to deliver efficient computer-supported tools for draft-ing the profile of each sherd and to automatically match it with the huge archives provided by available classifications (currently encoded only in drawings and written descriptions contained in books and publications). The system will also be able to support the production of archaeological documentation, including data on localisation provided by the mobile device (GPS). The platform will also allow to access tools and services able to enhance the analysis of archaeological resources, such as the open data publication of the pottery classification, or the data analysis and data visualisation of spatial distribution of a certain pottery typology, leading to a deeper interpretations of the past. Data analysis will be achieved as an ex-ploratory statistical analysis of data related to pottery. It will be mainly concerned with data about size, density, geo-localisation and chronology. The main objective of the exploratory analysis is to disclose statistical relationships (in statistical sense) between the different variables considered. Moreover, it will provide a comprehensive description of the available data, pointing out important features of the datasets, such as: where the information concentrates and where is missing, or where little data more would imply a relevant gain of information. There are different statistical techniques useful for exploratory data analysis, each one concentrating on particular aspects of the description we would like to give for the data. However, it is important to observe that the statistical techniques are not ex-ploratory as such, rather they are used in order to summarize main characteristics of data, identify outliers, trends, or patterns, i.e. they are used as explorative.

Concerning the analysis of pottery datasets, we will concentrate on the following tools:

- classification and clustering techniques, to be used for understanding whether or not some features of the data may possess convenient classifications in a number of categories/groups, subsequently suggesting meaningful interpretation of such categories;
- dimensionality reduction techniques, to be used in order to extract a small number of specific combination of features describing the greatest part of information and variability contained within the data. These specific combinations provide all at once a way to summarize data, and the identification of the major sources of variability;
- spatial statistics, point pattern analysis and Kriging methods will be mainly used in order to highlight the possible patterns within the spatial distribution of data;
- different predictive modelling techniques will be implemented mostly for suggesting where to look for more data in order to get relevant gain of information, or optimal strategies to perform testing.

The results of the data analysis will be made more understandable and easily explicable applying data visualisation techniques. Apart from the quantitative data analysis, data visualization is of extreme importance, in order to: provide an efficient way to understand a vast amount of data; allow non-technical people to do data-driven decision making; communicating the results of the data analysis (Llobera 2011). An important issue is the communicating the visual information about the relationships among different ceramic classes in the same location, the relationships between the location of the finding and the productive centre, and the relationships with pottery found in different locations. A web-based visualisation tools will be built following the principles of data visualization.

Following these guidelines, we will classify the different data into types (categorial, ordinal, interval, ratio types), and will determine which visual attributes (shape, orientation, colors, texture, size, position, length, area, volume) represent data types most effectively, so giving rise to the visualization, according to the basic principle of assigning most efficient attributes, such as position, length, slope, to the more quantitative data types, and less efficient attributes, like area, volume, or colors to ordinal or categorical types. The process of building the visualisation will be made interactive, letting the user associating the different variables with the different attributes, at the same time explaining the principles above. Moreover, the different relations within pottery production, trade flows, and social interactions, will be visualised applying the same principles, with graphs.

4. Conclusions

The possibilities of such system open to research actors, institutions and general public would be a dramatic change in the archaeological discipline as it is nowadays. Its impact on the field would dramatically change the profile of the professionals involved and will generate new markets.

References

Boyd D., Crawford K. (2012), Critical Questions for Big Data. Information, Communication and Society, (15), pp. 662–679

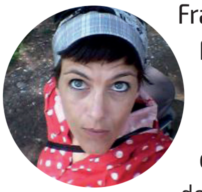
Gattiglia G. (2015), Think big about data: Archaeology and the Big Data challenge, Archäologische Informationen, (38), pp 113-124.

Llobera M. (2011), Archaeological Visualization: Towards an Archaeological Information Science (AISc), Journal of Archaeological Method and Theory, (18), p: 193–223.

Mayer-Schönberger V., Cukier K. (2013), Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, Boston, MA.

Samuelson P.A. (1954), The Pure Theory of Public Expenditure, Review of Economics and Statistics (36,4), pp 387-389.

Authors



Francesca Anichini - francesca.anichini@for.unipi.it

Francesca has worked as professional archaeologist for several years and directed excavations from Roman to Post-medieval ages. Since the 2010 she also works in MAPPALab at the University of Pisa as project and communication manager. She is one of developers of MOD - the Italian repository for Open Archaeological Data, and she deals with methodologic issues related to open data, archaeological potential and communication.

Gabriele Gattiglia - gabriele.gattiglia@for.unipi.it

He is a Researcher at the University of Pisa, coordinator of the ArchAIDE Project and leads the MAPPA Lab, which manages the MOD (Mappa Open Data), the Italian repository for Open Archaeological Data. He is devoted to digital application in archaeology, open data and Big Data issues in archaeology.