Big Archaeological Data.
The ArchAIDE project approach.

Gabriele Gattiglia, MAPPA Lab – Università di Pisa, gabriele.gattiglia@for.unipi.it
Francesca Anichini, MAPPA Lab – Università di Pisa, francesca.anichini@for.unipi.it

1. Big (archaeological) Data

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). In the scholarly world what constitutes Big Data varies significantly between disciplines, but it is possible to affirm that the shift in scale of data volume is evident in most disciplines, and that analysing large amounts of data holds the potential to revolutionise research, even in the Humanities. For a better understanding of the general concept of Big Data, it can be adopted the definition proposed by Boyd and Crawford (2012, 663): "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 and Cukier (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. Moreover, Big Data is about predictive modelling, i.e. about applying algorithms to huge quantities of data in order to infer probabilities, and it is about recognising the relationships within and among pieces of information. Furthermore, a Big Data approach is related to the information content of data: data are useful because they carry pieces of information, and they become information when they are processed and aggregated with other data. Finally, data are data because they describe a phenomenon in a quantified format so it can be tabulated and analysed, not because they are digital.

2. Datafication

Digitisation (i.e. the migration of pieces of information into digital formats) has changed archaeology deeply, and has increased exponentially the amount of data that could be processed, but digitisation, does not by itself involve datafication. Datafication is the act of transforming something (objects, processes, etc.) into a quantified format, so they can be tabulated and analysed (Mayer-Schönberger and Cukier 2013, 73); it promises to go significantly beyond digitisation, and to have an even more profound impact on archaeology. It can be argued that datafication puts more emphasis on the I (information) of IT, dis-embedding the knowledge associated with physical objects by decoupling them from the data associated with them (Gattiglia 2015). Datafication is manifest in a variety of forms and can also, but not always, be associated with sensors/actuators and with the Internet of Things. Moreover, 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.

3. 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 (www.archaide.eu) that aims to design, develop and assess 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 database built from the data provided by 2D printed catalogues and images. Starting from archaeologists needs, the project plan to deliver efficient computer-supported tools for drafting the profile of each sherd and to automatically match it with the huge archives provided by available classifications. The system will be able to support the production of archaeological documentation, including data on localisation provided by the mobile device (GPS), and will 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.

The first contribution of ArchAIDE is an as-automatic-as-possible procedure to transform the paper catalogues in a digital description, to be used as a data pool for an accurate search and retrieval process. On the other hand, the data collected through digitisation will be enriched by data collected by users during the recognition process. This will permit on-time data analysis and data visualisation. In fact, all the information encoded in the pottery identity cards (being them natively digital and including data on location, classification, dating, and so on) will be shared, visualised and integrated with cultural heritage information from different sources (archaeological repositories, Europeana, and so on). Real time comparisons between different archaeological sites and regions will be made possible, thus highlighting differences and commonalities in the economy of the ancient world.

Data analysis will be achieved as an exploratory 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 between the different variables considered, and to provide a comprehensive description of the available data. In fact, statistical techniques are used for summarizing 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, 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, for extracting 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, for highlighting the possible patterns within the spatial distribution of data;
- Predictive modelling, 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, A web-based visualization tool will improve accessibility to archaeological heritage, allow data-driven decision making, and communicate the results of the data analysis. Visualising 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 will generate new understanding about the dynamics of pottery production, trade flows, and social interactions. Visualisation tools will be built classifying the different data into types (categorial, ordinal, interval, ratio types), and determining which visual attributes (shape, orientation, colors, texture, size, position, length, area, volume) represent data types most effectively. The process of building the visualisation will be made interactive, letting the users associating the different variables with the different attributes. The possibilities of such system open to research and professional actors, institutions and general public would bring a dramatic change in the archaeological discipline as it is nowadays.

References

Boyd, D. and Crawford, K. 2012. “Critical Questions for Big Data. Information.” *Communication and Society* 15: 662–679

Gattiglia, G. 2015. “Think big about data: Archaeology and the Big Data challenge.” *Archäologische Informationen* 38: 113-124.

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