From digitization to datafication. A new challenge is approaching archaeology

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

Data, Big Data

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 (Royal Society, 2012, 7), even in the Humanities, on the other hand, in recent decades, archaeology has embraced digitisation. Moreover, the low cost and improvement in computing power (both software and hardware) gives the opportunity to easily aggregate huge amounts of data coming from different sources at high velocity: in brief we are in a Big Data era. Even if Big Data started in the world of Computer Science and are strongly connected to business, they are rapidly emerging in academic research, with scholars from different disciplines recognising the inherent research potential of analysing composite and heterogeneous datasets that dwarf in size and complexity those traditionally employed in their respective fields (Wesson and Cottier 2014; Gattiglia 2015). 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 scientific and scholarly world what constitutes Big Data varies significantly between disciplines, but we can certainly 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, producing hitherto impossible and unimaginable insights (Wesson and Cottier 2014, 1). For a better understanding of the general concept of Big Data, I adopt 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 (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. 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. Moreover, a Big Data approach is related to the information content of data. Data are useful because they carry pieces of information. As Clark's DIKW (Data Information Knowledge Wisdom) hierarchy (Clark 2004) and Hey's Knowledge Pyramid pointed out (Hey 2004), data are the building blocks of meaning, they are meaningless except for their relationship to other data. Data become information when they are processed and aggregated with other data, thereby we gain information from data when we make sense out of them (Anichini and Gattiglia 2015). Finally, we can say that data are data because they describe a phenomenon in a quantified format so it can be tabulated and analysed, not because they are digital.

Datafication

Digitisation has changed archaeology deeply. Digitisation usually refers to the migration of pieces of information into digital formats, for transmission, re-use and manipulation. Surely, this process has increased 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 is a new phenomenon brought out by the continuous development of IT technologies. 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. Datafication is the act of transforming something into a quantified format (Mayer-Schönberger and Cukier 2013, 73; O'Neil and Schutt 2013, 406). This is a key issue. As argued by (Cresswell 2014, 57) "two things that are making data suddenly big are the datafication of the individual and the geocoding of everything". To datafy means to transform objects, processes, etc. in a quantified format so they can be tabulated and analysed (Mayer-Schönberger and Cukier 2013). We can argue 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 (Bahga and Madisetti 2014, 37). 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.

ArchAIDE project

Introduction

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, a research unit of the Department of Civilisations and Form of Knowledge, 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. Our approach is driven by archaeologists needs; since we are aware of the caution of the discipline in front of the replacement of well-established methods, we plan to support this specific Humanities domain by exploiting what is already available in the Archaeology domain in terms of good practices and representation paradigms. We thus 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 (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. The integration of cultural heritage information from different sources, together with faster description, cataloguing and improved accessibility can be exploited to generate new knowledge around archaeological heritage. Data visualisation, for instance, would stimulate new research perspectives, and could enable new interpretation and understanding of history, and would bring archaeological storytelling to new audiences in a novel way. By means of a wider dissemination of user-generated content, the framework would permit to develop the culture of sharing cultural resources, research and knowledge.

From digitisation to datafication

The first contribution of ArchAIDE (www.archaide.eu) 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. This will entail: scanning (2D digitization) of the paper catalogue(s); segmentation and vectorialization of the graphical drawings proposed in those printed catalogues; and linking the graphical representation with the metadata reported in the catalogues. Since we are interested in designing automatic matching and retrieval features, digital description does not mean here only digitisation of the paper catalogues, but includes understanding the meaning of the graphic representation and its conversion in a format that includes shape (in vectorial format, not raster) and semantic. This process, naturally, will also require the definition of a semantically-rich digital vectorial representation for the pottery sherds and of each entire object able to represent not only the shape of the object, but also its subdivision in semantic components (e.g. rim, handle, foot, …). This representation, ideally, should be compliant with the existing representation, description and drawing standards used by archaeologists, to help both the digitisation phase (from “classical” documentation to digital) and the creation of the documentation (from digital back to “classical” documentation). A lightweight set of metadata (the subset considered crucial for the purposes of the project by our users and advisors, e.g. historical period, geographical region...) will be added to the extracted data. 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) in order to produce a really significant impact in the advancement of the discipline and in the accessibility for professional and non-professional users. 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. A web-based

visualization tool will improve accessibility to archaeological heritage and generate new understanding about the dynamics of pottery production, trade flows, and social interactions.

Data analysis will be carried on by the MAPPA Lab of the University of Pisa, and 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 (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 exploratory 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; Gattiglia 2015). 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, pionereed by (Bertin 2010, 83), and developed for instance in (Tufte 1990; Few 2006; Munzner 2014). 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.

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.

Bibliographic References

Anichini, Francesca and Gattiglia, Gabriele 2015. **"**Verso la rivoluzione. Dall'Open Access all'Open Data: la pubblicazione aperta in archeologia." *Post – Classical Archaeologies* 5: 299-326.

Bahga, Arshdeep and Madisetti, Vija 2014. *Internet of Things (a Hands-On Approach)*. Arshdeep Bahga Vijay Madisetti

Bertin, Jacques 2010. *Semiology of Graphics*. Esripress

Boyd, Danah and Crawford, Kate 2012. "Critical Questions for Big Data. Information." *Communication and Society* 15: 662–679

Clark, Donald 2004. "Understanding and Performance." Accessed November 15. http://www.nwlink.com/~donclark/performance/understanding.html%20.

Cresswell, Tim 2014. "Déjà vu all over again: Spatial Science, quantitative revolutions and the culture of numbers." *Dialogues in Human Geography* 4 (1): 54-58.

Few, Stephen 2006. *Information Dashboard design; The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.

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

Hey, Jonathan 2004. "The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link." Accessed November 15. http://inls151f14.web.unc.edu/files/2014/08/hey2004-DIKWchain.pdf

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

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

Munzner, Tamara 2014. *Visualization Analysis and design*. Boca Raton, FL: CRC Press.

O'Neil, Cathy and Schutt, Rachel 2013. *Doing Data Science*. Sebastopol, CA: O'Reilly Media.

Royal Society 2012. *Science as an Open Enterprise*. London, England: Royal Society.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36(4): 387-389.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Edward Tufte.

Wesson, Cameron B. and Cottier, John W. 2014. "Big Sites, Big Questions, Big Data, Big Problems: Scales of Investigation and Changing Perceptions of Archaeological Practice in the Southeastern United States". *Bulletin of the History of Archaeology* 24 (16): 1-11.