Think big about data: Archaeology and the Big Data challenge

Gabriele Gattiglia

Abstract – Usually defined as high volume, high velocity, and/or high variety data, Big Data permit us to learn things that we could not comprehend using smaller amounts of data, thanks to the empowerment provided by software, hardware and algorithms. This requires a novel archaeological approach: to use a lot of data; to accept messiness; to move from causation to correlation. Do the imperfections of archaeological data preclude this approach? Or are archaeological data perfect because they are messy and difficult to structure? Normally archaeology deals with the complexity of large datasets, fragmentary data, data from a variety of sources and disciplines, rarely in the same format or scale. If so, is archaeology ready to work more with data-driven research, to accept predictive and probabilistic techniques? Big Data inform, rather than explain, they expose patterns for archaeological interpretation, they are a resource and a tool: data mining, text mining, data visualisations, quantitative methods, image processing etc. can help us to understand complex archaeological information. Nonetheless, however seductive Big Data appear, we cannot ignore the problems, such as the risk of considering that data = truth, and intellectual property and ethical issues. Rather, we must adopt this technology with an appreciation of its power but also of its limitations.

Key words – Big Data, datafication, data-led research, correlation, predictive modelling

Zusammenfassung – Üblicherweise als Hochgeschwindigkeitsdaten (high volume, high velocity und/oder high variety data) bezeichnet, machen es Big Data möglich, dank dem Einsatz von Software, Hardware und Algorithmen historische Prozesse zu studieren, die man anhand kleinerer Datenmengen nicht verstehen kann. Big Data setzt einen neuen archäologischen Ansatz voraus: Die Bereitschaft, Massen von Daten zu nutzen, ungeordnete und heterogene Daten zu übernehmen, und Korrelation statt Kausalität zu akzeptieren. Kann die Unvollständigkeit archäologischer Daten einen solchen Ansatz verhindern? Oder sind archäologische Daten geradezu dafür prädestiniert, eben weil sie ungeordnet und unstrukturiert sind? Normalerweise handelt Archäologie mit großen und komplexen Mengen von Daten, oft fragmentarisch, und oft solchen, die aus verschiedenen Quellen und Disziplinen kommen und die selten im gleichen Format oder in der gleichen Skala vorliegen. Ist Archäologie bereit, mehr mit solchen Methoden zu arbeiten, die auf Daten basieren, und prädiktive und probabilistische Techniken zu akzeptieren? Big Data erklärt nicht, sondern informiert, bietet ein Modell für eine archäologische Interpretation an, ist eine Ressource und ein Werkzeug: Data Mining, Datenvisualisierung, Bildverarbeitung und quantitative Methoden können gemeinsam dazu beitragen, komplexe archäologische Informationen zu verstehen. So verführerisch Big Data auch sein mag, man sollte die Probleme nicht leugnen: Es besteht die Gefahr, Daten als absolute Wahrheit zu betrachten, zudem bestehen Fragen verbunden mit intellektuellen Rechten und Ethik. Wir können diese Technologie adaptieren, aber wir sollten ihre Stärken und Grenzen erkennen.

Schlagworte – Big Data, Datenerfassung, datengeführte Forschung, Korrelation, Vorhersagemodelle

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, namely we may realise innovative things by combining data in new ways. 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. In recent decades, Archaeology has embraced digitalization. This process has increased exponentially the amount of data that can be processed, but unfortunately, archaeological data are sometimes kept isolated in what we could call *data silos* or, with a more suggestive expression, *data tombs* to hint the fact that the data are buried and closed off from the rest of the archaeological community (Wren & Bateman, 2008). Many archaeologists seem to be unaware that the value of research data increases if they are available open access. The use of digital technologies is fostering the development of e-research that is the way scientific knowledge is produced and shared (Beaulieu & Wouters, 2009; Royal Society, 2012). Sharing has become a new scientific paradigm, and, if properly sustained by economic and political choices, will lead to open access to research data, making data openly available to public and private stakeholders, and to citizens (Wessels et al., 2014, 49). Moreover, the low cost and improvement in computing power (both software and hardware) gives us 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 & Cottier, 2014). In recent years, even archaeology is approaching the Big Data topic: in this paper I want to discuss what is the mea-

Received: 26 November 2014
accepted: 12 June 2015
published online: 23 July 2015

ning of the term Big Data, the pros and cons of Big Data, and if a Big Data approach can be applied to archaeology from both a theoretical and practical point of view.

Big Data: is there only one possible definition? Like many popular buzz-words, Big Data lacks consensus on a clear and consistent definition; Gantz and Reinsel (2011) define it as „A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis". Bloomberg (2013) adopt a simpler definition of Big Data as „a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques". In the ICT world, Big Data are usually defined within the Gartner glossary (2013) as high volume, high velocity, and/or high variety data, namely data „that demand cost-effective, innovative forms of information processing for enhanced insight and decision making", while in the scientific and scholarly world what constitutes Big Data varies significantly between disciplines. The Royal Society of Science (2012, 12) defines Big Data as unstructured data that require massive computing power to be processed, distinguishing them from Broad Data, namely structured data freely available through the web to everyone, and it argues that research data are generally not Big Data, and that they cannot be easily structured as Broad Data (Royal Society, 2012, 22). Evidently, the Royal Society of Science consider research data structured, but not open, data. Even so research data may need massive computing power to be processed. 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 & Cottier, 2014, 1). For a better understanding of the general concept of Big Data, I prefer to adopt a wider definition such as the one 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 cross-reference 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 us 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. Starting from this definition, we are able to outline four main theoretical aspects that stand behind a Big Data approach.

The first one is connected to the possibility to harness all the available data. As we discussed above, this is not intended in an absolute sense, but in a relative way: relative to the comprehensive dataset. The consequence of using the full set of data is that the concept of sampling loses the prominence it actually has. Taking advantage of all the available data (or at least as much as possible) makes it possible to illuminate the connections that are otherwise hidden in the abundance of data, and to look at the details, or to explore new ways of producing scientific knowledge. Effectively, ‚Big' does not mean only to understand a phenomenon on a wide scale, but also to analyse data in order to reach a level of granularity that samples cannot assess, because using all the data lets us see details we never could when we were limited to smaller quantities. From an archaeological point of view, if we analyse only the potsherds coming from sampled assemblages we would not able to have a complete pattern of the overall trade as when we use all the data available.

The second trait is related to the quality of data. Scholars strenuously defend the high quality of their data, and they are proud of their exactness. Unfortunately, if we decide to use a high volume of data of different and heterogeneous provenience, we cannot pretend to achieve the same level of accuracy, and we must accept messiness. Although Mayer-Schönberger and Cukier (2013) are certainly not suggesting that what works for Big Data applications is valid for all data applications, I disagree with the idea that accuracy is an obsession of the information-deprived era, but I agree with the fact that if we want to gain the benefit of working with Big Data, messiness is inevitable, for the reason that it is generated by adding more and more data, by combining different sources, by the inconsistency of formatting, and by the extraction and the transformation of data. Big Data forces us to re-think data quality, and to manage the quality/messiness question. First

of all, we have to consider data metrics: institutions need better ways of measuring the quality and impact of the data, for instance establishing practices for providing for peer review of data, including scientific data (Costas et al., 2013; Habert & Huc, 2010), as well as citing datasets in the same way as journal papers are currently cited in order to provide impact factors, and inaugurating forms of open peer review much as social media (Kansa & Whitcher Kansa, 2013; Piwowar et al.. 2007; Piwowar & Vision, 2013; Pöschl, 2010). For example, academics could afford the evaluation process of research data, and industry could contribute in ensuring quality of business and social media data (Wessels et al., 2014, 61). The problem of data quality has already emerged in crowdsourcing contexts (Saengkhattiya et al., 2012). In the Earth observation domain, for example, the need to include observations from unconventional and non-scientific sources, such as non-expert citizens, has stimulated solutions for the representation of data quality, such as encoding definitions of quality in metadata and data formats, and adding information through metadata enrichment and user's annotations (Wessels et al., 2014). Finally, we must be conscious of the fact that a lower data quality sometimes „enables bigger data-driven insights, which means that sometimes using a bigger amount of lower-quality data is better than using a smaller amount of higher-quality data" (Harris, 2013).

The third characteristic of 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 & Gattiglia, 2015). Moreover, 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. The act of transforming something into a quantified format is called datafication (Mayer-Schönberger & Cukier, 2013, 73; O'Neil & 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". Datafication promises to go significantly beyond digitalisation, and to have an even more profound impact, challenging the foundations of our established methods of measurement and providing new opportunities. Digitalisation 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. To datafy means to transform objects, processes, etc. in a quantified format so they can be tabulated and analysed (Mayer-Schönberger & 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 (Ericsson, 2014, 6). 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 & Madisetti, 2014, 37). Moreover, a key differentiating aspect between digitalisation and datafication is the one related to data analytics: digitalisation uses data analytics based on traditional sampling methods, while datafication fits a Big Data approach and relies on new forms of quantification and associated data mining techniques, which permit more sophisticated mathematical analyses to identify non-linear relationships among data, allowing us to use the information for massive predictive analyses.

The fourth (and last) characteristic of a Big Data approach is the most theoretical (and disputed) aspect of all. Will Big Data be the final chapter of Science as we know it? In other words, as Anderson (2008) suggests, will Big Data abolish models, theories, and hypotheses? Applying the Big Data paradigm means a shift from a more traditional hypothesis-driven approach, to an evidence-based data-driven approach (Brynjolfsson et al., 2011), able to produce less biased and more accurate outcomes. Data-led science does not represent the end of hypotheses and theory, but simply allowing „the numbers to speak for themselves" (Anderson, 2008), Big Data illuminates the correlations between data, making clear the patterns and offering us novel and invaluable insights. Correlations do not imply causation. In other words, a correlation between two variables does not necessarily imply that one causes the other, or to use a logic argumentation is not a sufficient circumstance. As suggested by Tufte (2004, 4) „observed covariation is a necessary but not sufficient condition for causality, (…) but it sure is a hint". Indeed, correlation is used to infer

causation; the important point is that such inferences are made after correlations are confirmed as real and all causational relationship are systematically explored (Aldrich, 1995; Bollier, 2010, 4; Pearl, 2009). This means that Big Data makes us renounce to the principle of causation, but not hypotheses and models: in a data-driven approach they come after and not before data analysis. At its core, a correlation between two data values (variables) measures the statistical relationship by which they are governed by common causes (Aldrich, 1995). If two variables are correlated that means that when one changes, the other is predicted to change as well. The potential of computer-aided visualisation of data, for instance, permits us to identify correlations and explore data as a way to develop and to test new models for extra investigation, and to validate hypotheses. Correlations offer pretty clear insights, that help us in capturing a phenomenon not by recognising its inner workings but by „identifying a useful proxy for it" (Mayer-Schönberger & Cukier, 2013); in this way they allow us to make predictions through the many mathematical and statistical methods we have to analyse relationships, and to demonstrate the strength of them with certainty. Big Data does not abolish theory and models; on the contrary, we can affirm that „Theory is about predicting what you haven't observed yet" (Bollier, 2010, 6).

All that glitters is not gold

‚Big Data' is rapidly becoming a research and scientific trend, thereby the number of scientific papers about Big Data increased faster per year than the best exponential curve, since the first appearance of the term in the 1970s (Halevi & Moed, 2012, 3-4). Unsurprisingly, the top scientific fields are Computer Science, Engineering, Mathematics, Business, and the Social and Decision Sciences, but there is also a growing interest in Big Data in the Humanities (Halevi & Moed, 2012, 4). Nonetheless, many scholars deprecate a topic that they consider ‚trendy', the utility of working with ever larger amounts of data, and with data whose quality they cannot control, arguing that opening up research data in a beneficial way requires a gradual approach. Finally, some researchers (Cresswell, 2014) denounce the attempt of using a Big Data approach in science as the foolish attempt to map the world in Borges's poem *On Exactitude in Science*, others (Barnes, 2013; Bollier, 2010; Boyd & Crawford, 2012; Tufekci, 2012; Royal Society, 2012, 23) raise the dystopian possibility of a Big (Brother) Data effect deriving from the insistence of ICT Corporations that once data sets become big enough, then there will be no more need for sampling, because data will closely match the world itself.

Even though, as discussed earlier, Big Data will not imply the end of theory or the end of data quality, Big Data definitely involve perils and problems. In this paper I will focus on what I deem the two main facets. The first one is data fetishism. Researchers can run the risk of considering data as truth, valuing them for „what they are rather than for what they do" (Barnes, 2013, 299). Although Big Data open up new research possibilities, without a proper contextualisation, an appropriate research design, and an information management plan to decide if a Big Data approach is useful, researchers may be overwhelmed by *data deluge* to the point of hindering their ability to address even ordinary research questions (Wesson & Cottier, 2014). No matter how comprehensive, or how sophisticated algorithms become, or how well analysed the data are, Big Data need to be accompanied by big judgment, that only researchers (and in our case archaeologists) can provide. This is the case, for example, with spurious correlations. The term describes misleading correlations which appear when the quantity of data increases by an order of magnitude, in which variables give the impression of being connected even though they are not (Aldrich, 1995).

The second facet concerns the ethical, legal and cultural issues surrounding the use of Big Data. The strength of a Big Data approach is to aggregate data from different disciplines, countries and researchers. A research project in landscape archaeology may be interested in collecting and analysing archaeological, geographical, environmental, palaeoenvironmental, geological, anthropological, climate, LiDAR, satellite, multispectral data, and also social media data and so on. This aspect, as well as open access to research data, raises major legal and ethical challenges, including considerations of intellectual property ownership, freedom of information, privacy laws, data protection laws, and cultural challenges. Some countries, for example, exclude data generated by governments from copyright, as well as information contained within databases; in other cases, a sui generis right provides legal protection for databases (Wessels et al., 2014). The key challenge regards the acceptance of high-level data-sharing principles, such as the full and open exchange of data and metadata made available with permissive copyright licenses (for instance Cre-

ative Commons BY or BY-SA, or Public Domain licenses), with minimum time delay, and either free of charge or with only the costs of reproduction (Anichini & Gattiglia, 2014; 2015).

Archaeology and the Big Data challenge

Although the origins of Big Data lie in Computer Science, Archaeology, as other disciplines, has been forced to meet the challenges of an era in which you „either go big or go home" (Wesson & Cottier, 2014, 1); thus the number of archaeologists involved in Big Data research is undoubtedly growing. Nevertheless, among archaeologists there is a lack of perception of being part of the Big Data world, and there seems to be a sense that combining archaeology and mathematics is somehow an ill fit (Newhard, 2013), so in this section I will review if archaeology is a suitable area of study for Big Data.

What constitutes the full set of data in archaeological research is often difficult to define. Harris (2006) and Lock and Molyneaux (2006) consider it a question of scale, while Wesson and Cottier (2014) propose a definition based on the spatial extension and the quantity of artefacts, suggesting that Big Data in archaeology is correlated to the dimension, „larger than those recovered in the majority of archaeological investigations", of the datasets resulting from large-scale, single-site excavations of more than a hectare, and multi-site investigations of corresponding spatial dimensions. In my opinion Big Data need not necessarily involve big archaeological interventions to yield big insights. As Leetaru (2012) argues, the full set of data in the Humanities concerns more the aggregation of big datasets, such as Wikipedia, and it is more a methodological approach, that does not depend on the spatial limits of archaeological investigations, but for example, on the aggregation of many of them. For instance, the majority of present-day archaeological interventions in Europe and in the United States are professional development-led investigations, generally of limited spatial extent, which produce low complexity assemblages (Wesson & Cottier, 2014). In this case, the Big Data approach is provided by the aggregation of both academic-based and commercial-based investigations of variable spatial extents, used, for instance, to study a wide-scale chronological question, or an archaeological landscape issue.

Aggregation of a wide variety of data is a key factor in Big Data. Archaeology perfectly fits this aspect. In fact, Archaeology has a long history of multidisciplinary research collaboration, to which, in the last decades with the coming of digitalisation, is added a systematic collection of data, that unfortunately produced mainly research questions at regional and culture-specific scales (Steckel, 2007, 18). Big Data develops this existing attitude to multidisciplinary research, and holds the potential to return transformative results with impacts cascading far beyond Archaeology, also strengthening the dissemination of research results (Kintigh et al., 2014; Steckel, 2007). We can suggest that the more the data from different disciplines are available, the better we can describe the general pattern of a phenomenon.

Archaeological data are messy and difficult to structure by definition: archaeological data structures are arbitrary, and there is no question about the interpretative character of their nature. Archaeological data structures, be they simple or complex, represent different ways of organising data, and are designed to achieve specific goals, such as to facilitate data retrieval, or to occupy a minimum of storage space. Moreover, data structures are fundamental for the application and development of algorithms (Llobera, 2011). Normally, archaeology deals with the complexity of large datasets, fragmentary data, data from a variety of sources and disciplines, rarely in the same format or scale. From this point of view, archaeological data do not preclude a Big Data approach, on the contrary Big Data is perfect for analysing them: „Archaeology is a place within the social sciences and Humanities where the nature of the work deals with Big Data" (Newhard, 2013). However, smaller scale data (e.g. data created for use by an individual researcher) often has poorer data modelling. These informal models may impede later data reuse and attempts to aggregate them at a higher scale (Kansa et al., 2014, 66). On the other hand, data reuse is a major challenge in Archaeology, as pointed out by Faniel et al. (2013), and it depends on the availability of data in useful forms too. At present, archaeological data require data cleaning and transformation procedures in order to be aggregated with other data. These data curation steps are common in Big Data, for example, in the case of social media data (Boyd & Crawford, 2012, 667), on the other hand Big Data will improve the data reuse experience and the standards development, in order to permit faster and less subjective analysis (Boyler, 2010, 13). Finally, with the general problems of data quality previously discussed, archaeological data cannot be used uncritically. It is necessary to find solu-

tions to manage data quality, establishing practices for providing open peer review of data, and encoding the evaluation of data quality through metadata enrichment and user annotation. Archaeologists are very capable of assessing their sources, and working with information collected by other researchers, even from different discipline; now they have to learn how to examine data. Thus, despite the considerable problems raised by data quality, digital techniques can mine existing archaeological collections to highlight anomalies, and to verify the quality of data. Big Data itself can be a powerful tool for providing data metrics (Lane, 2012).

Digitalisation has changed archaeology deeply. As already discussed, it has boosted the volume of data that can be analysed, but digitalisation does not involve datafication. As archaeologists, we are used to record information on paper, on computer, or on mobile devices, and we are well-aware that it is easier to create new datasets than transform old ones, because it takes energy and time to move information from analogue versions to digital ones. To datafy archaeology would mean to produce a constant flow of data, starting from the data produced by archaeological practice, such as locations and relationships between finds and sites. Besides, to datafy does not mean to record data and information more quickly in the field, but to record new information. Datafication represents a flow of data that the archaeological community should have available with minimum time delay, to process again and again. As Llobera (2011, 217) argues, these new data can modify the way we conduct our analyses, increase our capacity to process and visualise information in novel ways, and more decisively, provide new ways of doing archaeological research. This process requires a strong cultural and theoretical framework: changes have to be more qualitative than quantitative, and must involve theoretical orientations. From a cultural point of view any researcher must be aware of the opportunity of sharing data for improving their research; from a theoretical point of view, archaeological theory should shift towards data-driven research and a Big Data approach. Is archaeology ready to move towards data-led research, and to accept predictive and probabilistic techniques? In the last 20 years, predictive modelling has been used mainly as a decision-making tool in Cultural Resource Management (CRM), and less for the definition of site location or the interpretation of the spatial patterning of archaeological sites. The use of predictive models in CRM has produced both enthusiasm and criticism. Conversely, the recent practice of preventive archaeology shows that the use of predictive models in the early stages of land management planning is very successful for the protection of the archaeological heritage (Verhagen & Whitley, 2011). The use of predictive modelling in archaeology is connected with the rise of the New Archaeology in the late 1960's. By the 1980's, two primary lines of models were developed: models to identify spatial suitability, and models targeted to correlative statistical summaries that could be applied in unsurveyed areas (Gumerman, 1971; Verhagen & Whitley, 2011). The approaches used were mostly based on statistical modelling techniques, with a number of different methods based on regression, correlation, Bayesian statistics, and Kriging/coKriging models. As an alternative to statistical models, mathematical modelling has been applied. The latter has the advantage of allowing for the introduction of explicit working principles, by means of equations ruling the models. These equations contain additional information compared to statistical modelling, and include techniques like map algebra, trend surface analysis, cost distance models, Dempster-Schafer theory, and agent-based models (Drennan, 2010; Hodder & Orton, 1976; Kamermans et al., 2009; Wheatley & Gillings, 2002). For instance, Dubbini and Gattiglia (2014; 2013) used the relations-based PageRank algorithm (Langville & Meyer, 2006) to predict archaeological potential. On the other hand, statistical modelling is more indicated when no information at all is available about explicit working principles of the models. Archaeologists collect, organise, process, and synthesise data to investigate relationships and correlations so as to develop models and interpretations about environmental and human interactions, crossing across the disciplinary boundaries of the humanistic, social, natural, mathematic, and computer sciences (Newhard, 2013). Usually, archaeologists elaborate ‚reasoning artefacts' (Gooding, 2008; Thomas & Cook, 2005, 36), as an intermediate step between observations and interpretation, providing new explorations of correlations between data. The correlations are useful for archaeological interpretation, because archaeology, unlike the natural sciences, is further from the deterministic dualism of cause and effect. For this reason, Big Data approaches are effective on account of the fact that they inform, rather than explain, and that they expose patterns for archaeological interpretation, providing the opportunity to test new hypotheses at many levels of granularity. Data visualisation can provide an important contribution

to the comprehension of great amounts of data, and to make anomalies and correlations emerge. Unfortunately, as underlined by Llobera (2011, 213) data representation has not received as much attention as it should, especially in the light of the central role it has „in the production of knowledge and its potential to precipitate different interpretations". There is a strong conjunction between data and theory, a linkage that has not been exploited by archaeologists, and for this reason it has not produced new forms of data collection, representation and processing. The impact of computer applications in archaeology and FOSS (Free and Open-Source Software) in archaeology has been surprisingly limited; it has not been part of any radical change in how archaeology is done. Computer applications in archaeology have suffered from a deficiency of theory; they were unable to propose new developments, or new forms of conducting archaeological research, including new methods and standards of handling, processing and modelling information. This is related to the fact that computer applications are still marginal and reduced to a desirable technical skill, but there is insufficient awareness that the connection between computer application and archaeology provides new paradigms and/or research venues (Llobera, 2011). FOSS in archaeology seems on the point of losing this battle; will Big Data in archaeology lose the battle too? It should not, if it can overcome the absence of a proper academic curriculum. In other words, it is necessary to provide future archaeologists with a level of competency in both Archaeology and Computer Science, so as to enable them to move from one discipline to another with ease. Only proper training can permit archaeologists to participate in the development of new IT tools consistent with archaeological interests, and to foster a deeper conceptual understanding of how computer applications work as a necessary step towards the creation of new ones (Llobera, 2011; Lock, 2009). The full benefits of Big Data are only possible if such training is in place for archaeologists to gain the benefits themselves. There is a growing need for data archaeologists, namely researchers with skills for understanding the complexities of data, and abilities to synthesise and analyse information. Although the amount of data generated is growing exponentially, archaeologists are rarely included in the list of Big Data scientists, even if they have developed capacities to organise, manage, mine, and analyse large sets of data, and to extract meaning and insights. We need to educate more archaeologists with formal training in computational fields, since acquiring, organising, and analysing data are skills that should not be relegated to one single discipline. We need to overcome the concept of digital humanities or digital archaeology; today it should be expected that anyone leaving university can be assumed to have literacy in data mining and data processing.

Big Archaeological Data

It is not the objective of this paper to examine in detail the technological aspects of Big Data. The capability to gather huge amounts of data requires appropriate computer infrastructures, architectures, and analytical techniques (Boraso & Guenzi, 2013; Demchenko et al., 2013; Leetaru, 2012; Snow et al., 2006). A Big Data infrastructure needs to support data management operations and processes, administering access to data and data security services to researchers. A Big Data architecture framework must include the following components: one element reserved to data models, structures and types (data formats, file systems, etc.); one dedicated to Big Data management (Big Data lifecycle, transformation/staging, archiving); one for Big Data analytics (Big Data applications, presentation, visualisation); one for Big Data infrastructure (storage, computing, network, Big Data operational support), and one for Big Data security. The component addressed to data models and structures manages the raw data, the structured data and datasets that went through data filtering and processing, the published data that support research results, and data linked to publications, as described by the European Commission (2012) report. The same element handles metadata and all the information about the processes involved in the transformation of data. As already discussed, each stage of this transformation process needs different data structures, models and formats (data described via a formal data model, data described via a formalised grammar, data described via a standard format, arbitrary textual or binary data) including also the opportunity to process both structured and unstructured data (Demchenko et al., 2013). The Big Data analytics infrastructure is the place where the Big Data applications are supported. One of the best solutions is to base it on the Hadoop framework (Hadoop related services and tools; specialist data analytics tools; databases/servers SQL, NoSQL; Massively Parallel Processing databases) that can be easily integrated with analysis software like R, Apache Solr and many

others. These software are free and open source, and can be installed on commodity hardware; on the other hand, Big Data analytics tools are currently offered by the major cloud services providers such as: Amazon Elastic MapReduce and Dynamo, Microsoft Azure HDInsight, and IBM Big Data Analytics.

To set up this kind of infrastructure requires financial investment, so it becomes worthwhile only if we have the capacity to investigate big archaeological questions, specifically challenges that require data from varied disciplines and at different scales, and that address present-day problems. I will try to demonstrate how some of these issues are perfectly suitable for a Big Data approach. A topic such as the emergence, persistence, evolution and failure of market systems requires a great volume of archaeological and historical data about short-term fluctuations in production, supply, value, price and consumption to investigate a theme that is central to the advent of the modern world system. Issues connected to resilience, persistence, transformation, and collapse necessitate high volumes of data coming from archaeology, and the social and natural sciences, and related to a wide range of societies. Such research needs to analyse population, productivity, and climate data at different scales. On the other hand, integrating insights from ecology and archaeology can contribute to our present-day understanding of the role of diversity and complexity in the resilience of socioecological systems. Better awareness of the correlations between diversity and complexity at different scales can inform contemporary policies dealing with sustainability. Bearing in mind the increasing concern about the sustainability of demographic and environmental trends and pressure, few issues are more crucial than the possibility that our planet cannot support continued population growth and accelerated use of natural resources. The relationships between environment, population, settlement and mobility can also be studied using a Big Data approach, because biological, environmental, sociological, historical, anthropological and archaeological data of varied spatial and chronological scales need to be aggregated. Even the response to sudden environmental modification requires the integration of data from archaeology, zooarchaeology, paleoecology, sedimentology, seismology, and geomorphology. Considering how present-day migrations are often associated with drought, floods, warfare, political unrest, and religious persecution, this is a topical issue.

Starting from the aforementioned considerations, the MAPPA Laboratory of the University of Pisa, together with the author, is planning a research project that we decided to call the Big Archaeological Data Project (the BAD Project). This will examine in depth the theoretical aspects of a Big Data approach in archaeological research before moving to Big Data analysis of three different aspects: predictive modeling; the perception of archaeology; and historical/archaeological analysis. In the first case, the project represents the improvement of the urban PageRank model of archaeological potential elaborated during the MAPPA Project (Anichini et al., 2012; Anichini et al., 2013; Dubbini & Gattiglia, 2013). The MAPPA project itself can be considered to be an in *nuce* Big Data project for the use of high variety data, mathematical applications, predictions, datafication of urban archaeology, and open access to research data. The BAD project will enhance this model to fit a larger spatial scale, applying the Big Data paradigms discussed above. For the historical/archaeological analysis our idea is connected to another MAPPA project issue, that of urbanism and urban landscape, enlarging the scale to the European context (Bettencourt et al., 2008; Brogiolo, 2011; Cowgill, 2004; Gattiglia, 2014; Lilley, 2009; Marcus & Sabloff, 2008; Smith, 2010). Cities are the origin of present-day society and their role in both social and economic life is growing. To study the expansion of cities in the past, as well as their problems, difficulties, and collapse, can help us to understand the directions in which the urban centers of the present will develop. Urban landscape studies need to incorporate high volumes of varied data at different spatial scales. Historical cities provide archaeological, historical, social, demographic, palaeoenvironmental, and geomorphological data that illuminate the layout, organization, and data visualisation of urban life. Archaeological data on cities range from small finds to the patterns of urban fabric covering great spatial extents and presenting large chronological depth. Consequently, characterising long-term urban fabrics and animating associated behaviours via computational modelling requires a high volume of data and substantial computational infrastructure.

Sentiment analysis will also be applied to an international archaeological site to test public perceptions of archaeology. Sentiment analysis and opinion mining techniques, including the analysis of feelings and the detection of user opinions, have gained increasing interest in the context of the analysis of social data extracted from social networks, review sites, blogs, etc. (Pang & Lee,

2008). Typically, the techniques used are based on a classification of terms and adjectives according to positive, negative, or objective characteristics (Baccianella et al., 2010); moreover, using textual analysis, it is possible to derive higher level characteristics from these terms (Martineau & Finin, 2009), to be used as classifiers (e.g. Support Vector Machine) for the analysis of portions of text (Mullen & Collier, 2004), or for considering how the meaning of these terms changes on the basis of related terms (Taboada et al., 2011). Content-based recommender systems, i.e. information filtering system that seek to predict the rating or preference that user would give to an item (Ricci et al., 2011) can be applied in a number of cases of interest to this project, such as web pages, news and events, restaurants, multimedia data, museums, monuments, and works of art (Pazzani, 2007), and can be extended to include social information (Balby Marinho et al., 2012). In order to solve the problem of large-scale recommendations within noisy and scattered data (as typically happens when user preferences are given on a voluntary basis on social networks) methods of matrix factorisation can be applied (Koren et al., 2009), and scalable solutions may be defined to ensure adequate performances (Takacs et al., 2009).

Conclusion

Big Data is a new technological trend in science, but there are not yet many academic papers related to Big Data, and in most cases they are focused on some particular technology or solution that reflects only a small part of the whole argument. The same applies to the definition of Big Data, for which there is not a well-established terminology. Thus, this is the right moment for Archaeology to choose its main road for a Big Data approach. In this paper I demonstrated that Big Data is a solution to resolve existing archaeological challenges, and that Big Data is more a methodological approach than a question of high volume, high velocity, high variety data. Although it can be defined also as high value, and high veracity, a Big Data approach is better characterised by its paradigm: Big Data as All Data, namely the opportunity of using all data available, messiness/quality, datafication, and correlation/prediction. The use of Big Data does not imply the end of archaeological theory, or even the end of archaeologists: no matter how comprehensive or well analysed the data are, they need to be complemented by big judgment. As much as there is a need for skills in data management and manipulation, there is a need for understanding what the data mean. These skills must not be delegated to data scientists, because the skills in application, creativity, and synthesis are equally developed in the Humanities. On the other end, without a sharing attitude among researchers is very difficult to apply a Big Data approach. For this reason we do not have to talk about Big Data, but we have to dream about Big Open Data.

Big Data will not mean the end of small-scale archaeological research; they will continue to make their own contribution to our understanding of the past. Big Data, however, as suggested by many scholars (Anichini & Gattiglia, 2014; Horsley et al., 2014; Kintigh, 2006; Kintigh et al., 2014; Llobera, 2011; Lock, 2009; Snow et al., 2006; Steckel, 2007; Wesson & Cottier, 2014) can radically transform archaeological practice, fostering new research questions, novel data visualization techniques, new competences, and an enhanced ability to address those big questions only archaeological research is capable of investigating. Perhaps, in the near future, Big Data could even enable us to count at least one major scientific research question that could be addressed through the use of archaeological data among those considered as the ‚25 most important questions in science' (Science, 2005).

References

Anichini, F., Fabiani, F., Gattiglia, G. & Gualandi, M.L. (2012). *Mappa. Methodology Applied to Archaeological Potential Predictivity 1.* Roma, Italy: Edizioni Nuova Cultura.

Anichini, F., Dubbini, N., Fabiani, F., Gattiglia, G. & Gualandi, M.L. (2013). *Mappa. Metodologie Applicate alla Predittività del Potenziale Archaeologico 2.* Roma, Italy: Edizioni Nuova Cultura.

Anichini, F. & Gattiglia, G. (2014). Verso un'archeologia 2.0. *Scienza & Società – Open Science Open Data. La scienza trasparente 17-18*, 103-114.

Anichini, F. & Gattiglia, G. (2015). Verso la rivoluzione. Dall'Open Access all'Open Data: la pubblicazione aperta in archeologia. *Post – Classical Archaeologies 5*, 299-326.

Aldrich, J. (1995). Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science 10(4)*, 364-376.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine 16 (07), July 23rd 2008*.

Baccianella, S., Esuli, A. & Sebastiani, F. (2008). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, 2200–2204.

Bahga, A. & Madisetti, V. (2014). *Internet of Things (a Hands-On Approach)*. Arshdeep Bahga & Vijay Madisetti.

Balby Marinho, L., Hotho, A., Jäschke, R., Nanopoulos, A., Rendle, S., Schmidt-Thieme, L., Stumme, G. & Symeonidis, P. (2012). *Recommender Systems for Social Tagging Systems*. New York, NY: Springer.

Barnes, T. (2013). Big Data, Little History. *Dialogues in Human Geography 3*, 297–302.

Beaulieu, A. & Wouters, P. (2009). E-research as intervention. In N. Jankowski (ed.), *E-research: transformation in Scholarly Practice*, (pp. 54 – 69). New York, NY: Routledge.

Bettencourt, L. M. A., Lobo, J. & Geoffrey, B. W. (2008). Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics. *European Physical Journal B 63*, 285–293.

Bloomberg, J. (2013). *The Big Data Long Tail*. http://www.devx.com/blog/the-big-data-long-tail.html [15.11.2014].

Bollier D. (2010). *The Promise and peril of Big Data*. Washington, DC: The Aspen Institute.

Boraso, R. & Guenzi D. (2013). Architetture scalabili per memorizzazione, analisi condivisione epubblicazioe di grosse moli di dati. In M. Serlorenzi (ed.). *ARCHEOFOSS. Free, Libre and Open Source Software e Open Format nei processi di ricerca archeologica. Archeologia e Calcolatori (supp. 4)*, 139-146.

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

Brogiolo, G. P. (2011). *Le origini della città medievale*. Mantova, Italy: SAP.

Brynjolfsson, E., Hitt, L. & Heekyung, K. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* http://ssrn.com/abstract=1819486 [15.11.2014].

Clark, D. (2004). *Understanding and Performance*. http://www.nwlink.com/~donclark/performance/understanding.html [12.3.2015].

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

Costas, R., Meijer, I., Zahedi, Z. & Wouters, P. (2013). *The Value of Research Data – Metrics for datasets from a cultural and technical point of view*. http://knowledge-exchange.info/datametrics [12.3.2015].

Cowgill, G. L. (2004). Origins and Development of Urbanism: Archaeological Perspectives. *Annual Review of Anthropology 33*, 525–542.

Demchenko, Y. & Ngo, C., Membrey, P. (2013). *Architecture Framework and Components for the Big Data Ecosystem* (Draft Version 0.2. 12 September 2013). http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf [15.11.2014].

Drennan, R. D. (2010). *Statistics for archaeologists*. New York, NY: Springer.

Dubbini, N. & Gattiglia, G. (2014). Mathematical models for the determination of archaeological potential. In G. Earl, T. Sly, T. Chrysanthi, P. Murreta-Flores, C. Papadopoulos, I. Romanowska & D. Wheatley (eds.), *Archaeology in the Digital Era. Volume II. e-Papers from the 40th Annual Conference of Computer Applications and Quantitative Method in Archaeology* (pp. 710-719). Amsterdam, Netherland: University Press.

Dubbini, N. & Gattiglia, G. (2013). A PageRank based predictive model for the estimation of the archaeological potential of an urban area. In A. C. Addison, G. Guidi, L. De Luca & S. Pescarin (eds.), *Proceedings of the 2013 Digital Heritage International Congress. Marseille*, 28 October – 1 November 2013 (pp. 571-578). Denver, CO: IEEE.

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J. & Yakel, E. (2013). The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. *JCDL 2013 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp.295-304). New York, NY: ACM.

Ericsson Telefonaktiebolaget LM (2014). *The Impact of Datafication on strategic landscapes*. http://www.ericsson.com/res/docs/2014/the-impact-of-datafication-on-strategic-landscapes.pdf [15.11.2014].

European Commission (2012). *Advancing Technologies and Federating Communities. A Study on Authentication and Authorisation Platforms for Scientific Resources in Europe. Final Report*. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf [15.11.2014].

Gattiglia, G. (2014). *Pisa in the Middle Ages: archaeology, spatial analysis and predictive modelling*. Roma, Italy: Edizioni Nuova Cultura.

Gantz, J. & Reinsel, D. (2011). *Extracting Value from Chaos*. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf [15.11.2014].

Gartner Glossary (2013). *Big Data definition*. http://www.gartner.com/it-glossary/big-data/ [15.11.2014].

Gooding, D. C. (2008). Envisioning explanation: The art in science. In B. Frischer & A. Dakouri-Hild (eds.), *Beyond illustration: 2d and 3d Digital Technologies as Tools for Discovery in Archaeology* (pp. 45–74). Oxford, England: Archaeopress.

Gumerman, G. J. (1971). *The Distribution of Prehistoric Population Aggregates*. Prescott, AR: College Press.

Habert, B. & Huc, C. (2010). Building together digital archives for research in social sciences and humanities. *Social Science Information 49 (3)*, 415- 443.

Halevi, G. & Moed, H. (2012). The Evolution of Big Data as a Research and Scientific Topic. *Research Trends 30*, 3–6.

Harris. J. (2013). *The Eight Law of Data Quality. The Data Roundable*. http://blogs.sas.com/content/datamanagement/2013/06/19/the-eighth-law-of-data-quality/ [15.11.2014]

Harris, T. (2006). Scale as Artifact: GIS, Ecological Fallacy, and Archaeological Analysis. In G. Lock & B. L. Molyneaux (eds.), *Confronting Scale in Archaeology* (pp. 39–53). New York, NY: Springer.

Hey, J. (2004). *The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link*. http://www.dataschemata.com/uploads/7/4/8/7/7487334/dikwchain.pdf [12.3.2015].

Hodder, I. & Orton, C. (1976). *Spatial analysis in archaeology*. Cambridge, England: University Press.

Horsley, T., Wright, A. & Barrier, C. (2014). Prospecting for New Questions to Define Anthropological Research Objectives and Inform Excavation Strategies at Monumental Sites. *Archaeological Prospection 21*, 75–86.

Kamermans, H., Van Leusen, M. & Verhagen P. (eds.) (2009). *Archaeological prediction and risk management*. Leiden, Netherland: University Press.

Kansa, E. C. & Whitcher Kansa, S. (2013). We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies 1*, 88-97.

Kansa, E. C., Whitcher Kansa, S. and Arbuckle B. (2014). Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology. *International Journal of Digital Curation 9 (1)*, 57–70.

Kintigh, K. (2006). The Promise and Challenge of Archaeological Data Integration. *American Antiquity 71*, 567–578.

Kintigh, K. W., Altschul, J. H., Beaudry, M. C., Drennan, R. D., Kinzig, A. P., Kohler, T. A., Limp, W. F., Maschner, H. D. G., Michener, W. K., Pauketat, T. R., Peregrine, P., Sabloff, J. A., Wilkinson, T. J., Wright, H. T. & Zeder M. A. (2014). Grand Challenges for Archaeology. *American Antiquity 79(1)*, 5–24.

Koren, Y., Bell, R. & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer 42*, 30 –37.

Lane, J. (2012). Big Data. *Research Trends 30*, 7–10.

Langville, A. N. & Meyer, C. D. (2006). *Google's PageRank and Beyond*. Princeton, NJ: University Press.

Leetaru, K. (2012). A Big Data Approach to the Humanities, Arts, and Social Sciences. *Research Trends 30*, 17–30.

Lilley, K. D. (2009). *City and Cosmos: The Medieval World in Urban Form*. London, England: Reaktion.

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

Lock, G. (2009). Archaeological Computing Then and Now: Theory and Practice, Intentions and Tensions. *Archaeologia e Calcolatori 20*, 75–84.

Lock, G. & Molyneaux, B. (2006). Introduction: Confronting Scale. In G. Lock & B. L. Molyneaux (eds.). *Confronting Scale in Archaeology: Issues of Theory and Practice* (pp. 1–11). New York, NY: Springer.

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

Marcus, J. & Sabloff, J. (eds.) (2008). *The Ancient City: New Perspectives on Urbanism in the Old and New World*. Santa Fe, NM: SAR Press.

Martineau, J. & Finin, T. (2009). Delta tfidf: An improved feature space for sentiment analysis. In: *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*, 258–261.

Mullen, T. & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In: *Proceedings of EMNLP, vol. 4*, 412–418.

Newhard, J. (2013). *Archaeology, Humanities, and Data Science*. http://blogs.cofc.edu/thearchaeoinformant/2013/08/01/archaeology-humanities-and-data-science/ [15.11.2014].

O'Neil, C. & Schutt, R. (2013). *Doing Data Science*. Sebastopol, CA: O'Reilly Media.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2 (1-2)*, 1–135.

Pazzani, M. & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (eds.), *The Adaptive Web*. LNCS (pp. 325–341). New York, NY: Springer.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge, England: University Press.

Piwowar, H. A., Day, R. S. & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increasead Citation Rate, *PLos ONE 2 (3)*, e308.

Piwowar, H. & Vision, T. J.(2013). Data reuse and the open data citation advantage. *PeerJ 1*, e175.

Pöschl, U. (2010). Interactive open access publishing and public peer review: the effectiveness of transparency and self-regulation in scientific quality assurance. *International Federation of Library Associations and Institutions Journal 36 (1)*, 40-46.

Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B. (eds.) (2011). *Recommender Systems Handbook*. New York, NY: Springer.

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

Saengkhattiya, M., Sevandersson, M. & Vallejo, U. (2012). *Quality in crowdsourcing: How software quality is ensured in software crowdsourcing* (Master dissertations, University of Lund). http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3168789&fileOId=3168790 [15.11.2014].

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

Science (2005): July 1.

Smith, M. E. (2010). Sprawl, Squatters, and Sustainable Cities: Can Archaeological Data Shed Light on Modern Urban Issues? *Cambridge Archaeological Journal 20*, 229–253.

Snow, D., Gahegan, M., Giles, C., Hirth, K., Milner, G. & Mitra, P. (2006). Information Science Enhanced: Cybertools and Archaeology. *Science 311*, 958–959.

Steckel, R. (2007). Big Social Science History. *Social Science History 31*, 1–33.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics 37 (2)*, 267-307.

Takács, G., Pilászy, I., Németh, B. & Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. *Journal of Machine Learning Research 10*, 623–656.

Thomas, J., & Cook, K.A. (Eds.) (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press.

Tufekci, Z. (2012). *Data Dystopia*. http://www.technologyreview.com/notebook/428210/data-dystopia/ [15.11.2014].

Tufte, E. (2004). *The Cognitive Style of PowerPoint*. Cheshire, CO: Graphic Press LLC.

Verhagen, P. & Whitley, T. G. (2011). Integrating Archaeological Theory and Predictive Modelling: a Live Report from the Scene. *Journal of Archaeological Method and Theory 12 (1)*, 49-100.

Wessels, B., Finn, R. L., Linde, P., Mazzetti, P., Nativi, S., Riley, S., Smallwood, R., Taylor, M. J., Tsoukala, V., Wadhwa, K. & Wyatt, S. (2014). Issues in the development of open access to research data. *Prometheus: Critical Studies in Innovation 32(1)*, 49-66.

Wesson, C. B. & Cottier, J. 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.

Wheatley, D. & Gillings, M. (2002). *Spatial Technology and Archaeology*. London, England: Taylor and Francis.

Wren, J. D. & Bateman A. (2008). Databases, data tombs and dust in the wind. *Bioinformatics 24 (19)*, 2127-2128.

About the author

Archaeologist, working at the MAPPA Lab of the University of Pisa (www.mappaproject.org). He shares his time between archaeological research and his family. He has written many articles, two books on medieval Pisa and is Scientific Director of the research project ‚Medieval Versilia'. Recently, being part of the MAPPA Project team, he started to work on mathematical applications in archaeology (with the mathematicians Nevio Dubbini), especially on predictive models, and on open data issue (with Francesca Anichini). He firmly believes that the sharing of archaeological data is a necessary step for the development of the discipline, for this reason is one of the creators and curators of Italian open data repository MOD (MAPPA Open Data). He is a member of the Editorial Board of the Journal of Open Archaeological Data, and collaborates with the Open Pompeii project for the open access to archaeological data of the ancient city. He actively supports the Italian National Association of Archaeologists (ANA), as a member of the Scientific Committee. You can find him on twitter (@g_gattiglia) and on academia.edu (https://pisa.academia.edu/GabrieleGattiglia).

Gabriele Gattiglia Ph. D.
Post doc Research Fellow
MAPPA Lab – Dipartimento di Civiltà e Forme del Sapere
University of Pisa
Via Trieste 38, 56126, Pisa, Italy
g.gattiglia@for.unipi.it