e-ISSN 2724-3923

magazén
Vol. 1 – Num. 1 – June 2020

From Digital Archaeology to Data-Centric Archaeological Research

Franco Niccolucci
PIN scrl – Servizi Didattici e Scientifici per l'Università di Firenze, Italia

Abstract Since the end of the 20th century the widespread use of digital applications in archaeology has legitimized their inclusion in the archaeological toolbox. Together with archaeological sciences, databases, GIS and other computer-based methods are nowadays present in every respectable archaeological investigation. This makes archaeology a peculiar discipline, where the scientific method combines with the historical one to produce new knowledge. However, the large availability of archaeological data creates the risk of a data deluge and may suggest using online information just to collect previous interpretations rather than to re-use the data supporting them. A 'Grand Challenges' list compiled some years ago includes important research questions that undergird contemporary issues and require an appropriate digital methodology to be addressed. The present paper discusses the benefits, or better the absolute need, of a data-centric methodology to address large-scale research. It argues that an acritical use of the so-called 'Big Data' approach may be questionable. It suggests how the combination of artificial intelligence with human intelligence is the key to progress into the understanding of phenomena of paramount societal importance for researchers and for the public at large.

Keywords Archaeological data. Data-centric research. Semantics. Archaeological ontologies. Big data.

Edizioni Ca'Foscari

Peer review

Submitted	2020-07-17
Accepted	2020-10-06
Published	aaaa-mm-dd

Open access

Citation Niccolucci, F. (2020). "From Digital Archaeology to Data-Centric Archaeological Research". *magazén*, 1(1), 35-54.

DOI 10.30687/mag/2724-3923/2020/01/002

1 Introduction

In a paper published about 20 years ago (Gardin 1999, 63), Jean-Claude Gardin stated that:

> un des problèmes majeurs de notre temps en matière d'information scientifique est le déséquilibre qui s'est instauré entre la quantité croissante des travaux publiés à l'intention des chercheurs que nous sommes, dans quelque domaine que ce soit, et le temps à peu près inchangé que nous pouvons consacrer à les lire.[1]

As Gardin mentions in the same article, a similar statement had been made almost 10 years before, in 1991, by Sir Anthony Kenny, then President of the British Academy, who declared that he could not hope to read more than a very small part of the articles published in the UK and the USA relevant for his reportedly *narrow field of interest*. In sum, the question of 'information deluge', i.e. the overwhelming quantity of data pertaining to the same subject is not new, and, according to Gardin, the development of information technology did not appear to have solved it in those early years:

> les nouvelles technologies de l'information ne répondent pas pleinement à la crise présente de l'information scientifique. (66)[2]

A few years ago, Keith Kintigh and several other US archaeologists, authors of a 2014 paper (Kintigh et al. 2014, 19) on the Grand Challenges for Archaeology, stated that:

> both the modelling and the synthetic research will require far more comprehensive online access to thoroughly documented primary research data and to unpublished reports and other documents detailing the contextual information essential for the comparative analyses.

In conclusion, the amount of information available online is exponentially increasing due to the improved availability of storage and to policies fostering openness of research results. This is a significant achievement, but without a solution, available information risks to become unmanageable because, as Gardin stated, the time available

1 "One of the major current problems concerning scientific information is the existing imbalance between the increasing quantity of works published by researchers like us, in whatever domain, and the almost unchanged time that we can spend on reading them" (this translation and the following ones are by the Author).

2 "New information technologies do not *fully* answer to the present crisis of scientific information" (Author's italics).

for reading has remained the same notwithstanding such increase.

Policies concerning Open Access, such as the FAIR (Findable-Accessible-Interoperable-Reusable) data principles (Wilkinson 2016, 2019) will help little in the retrieval of archaeological information unless accompanied by implementation guidelines and supported by effective software tools and services. The FAIR principles concern generic research data, and it is relatively easy to document scientific datasets with appropriate metadata in order to be able to fulfil such principles.[3]

On the contrary, for archaeological datasets, plain compliance with the FAIR principles is a significant step forward, but not enough to address the issues described above. This is due to the complex nature of archaeological datasets, where the scientific method combines with the historical one to produce new knowledge and the contribution of many different disciplines including physics, chemistry, materials science and biology adds to direct observation and to digital services like databases, GIS (Geographical Information Systems) and more. The issues discussed in the present paper concern, in particular, unpublished archaeological reports (the so-called 'grey literature'), which are the most difficult to manage, more than publications, for which there exist well-organised digital libraries. In what follows, reference to archaeological text documentation should be intended to include both.

In the present paper, potential advancements to survive to the archaeological data deluge are discussed. The first part will address data accessibility and interoperability, together with what might be called 'first level' findability. To implement full re-usability that requires 'advanced findability', more sophisticated technology is required. A discussion of available methods and tools to achieve it is presented in the second part of this article, including some caveats about potentially misleading shortcuts.

2 Gardin's Logicism

Before proceeding, Jean-Claude Gardin must be cited again. He must be credited for proposing 'Logicism' (Gardin 1980) in the last quarter of the twentieth century, an innovative approach on how IT can support archaeological research. The logicist proposal was criticised arguing that it proposed to simulate archaeological reasoning, with what appeared to critics as a lack of interest in the content.

3 Henceforth, when one of the four terms forming the FAIR acronym is mentioned with reference to the principles, it will be capitalized: e.g. 'Find' refers to the first of the FAIR principles while 'find' just means discover, retrieve, as usual.

Even in France, the logicist approach did not become a blockbuster and did not achieve a widespread acceptance, besides an attempt to use multimedia for the purpose (Gardin, Roux 2004).

Logicism had little fortune outside of France also because it did not belong to the Anglo-centric theoretical and methodological discussion. In Italy, besides the pioneering journal *Archeologia e Calcolatori*, computer applications at the time were not appreciated in the archaeological circles, with the notable exception of a handful of far-sighted distinguished scholars. The few researchers interested in such applications were more attracted by English and American models such as those, for example, presented at the annual CAA (Computer Applications and Quantitative Methods in Archaeology) international conferences (D'Andrea, Niccolucci 2000). Thus, also in Italy this methodology did not attract many supporters and had no application.

Regardless of any theoretical evaluation of logicism, a major obstacle to its diffusion consisted also in the need to adopt a novel system for the documentation, risking that all the accumulated results might become incompatible with it. But, on this regard, a French team has recently used a logicist approach to document the excavations in the church of Rigny with interesting results (Buard et al. 2019; Marlet et al. 2019; Zadora-Rio et al. 2020). This methodology provides interoperability with other systems by using for its concepts and its inference chain (the reasoning) the standard ontology used in Cultural Heritage with the appropriate archaeological extensions, described in the next sections. This could make the logicist approach interoperable with the archaeological documentation standards, keeping the richness of documentation it provides and overcoming the objection of being an 'alien' in a world of databases, archaeological GIS and systems based on semantics. This work is currently progressing, as shown in a very recent methodological paper (Nuninger et al. 2020).

3 Data Availability and Access

Back to Kintigh's statement mentioned above, several initiatives started collecting and making available online archaeological documentation. The most important initiatives to store, preserve and make archaeological data available online are ADS (Archaeological Data Service) (Richards 1997, 2017) in Europe, led by Julian D. Richards at the University of York. ADS is a UK repository for digital archaeological records existing since 1996, storing a large number of unpublished reports. In the USA, tDAR (the Digital Archaeological Record) aims at a similar target (Kintigh 2006; McManamon, Kintigh, Brin 2010). tDAR is led by Keith W. Kintigh at Arizona State University. There are also many specialised databases such as, among others,

ROAD (ROCEEH Out of Africa Database) created by the ROCEEH[4] (The Role of Culture in Early Expansions of Humans) project led by the University of Tubingen and funded by the Heidelberg Akademie der Wissenshaften (Heidelberg Academy of Sciences), dedicated to palaeoanthropological and palaeoenvironmental data; and Open Context,[5] a US initiative publishing research data at a fee.

Since 2013, the ARIADNE (Archaeological Research Infrastructure for Archaeological Data Networking in Europe) project is an EU-funded integrating activity to aggregate archaeological datasets in Europe (Niccolucci, Richards 2013; Meghini et al. 2017; Aloia et al. 2017; Niccolucci 2018). It has created and manages a registry of about 2,000,000 archaeological datasets, searchable according to facets such as time, place and object type. The ARIADNE extension, ARIADNEplus (Richards, Niccolucci 2019; Niccolucci, Richards 2019), also an EU-funded project, recently started an ambitious plan to extend its coverage both geographically and thematically, using state-of-the-art digital technology to support searching and finding. ARIADNEplus is fully compliant with the FAIR data and Open Science principles.

Both ARIADNE and ARIADNEplus work as aggregators. They collect metadata from organisations managing a data repository such as ADS or other institutions in Europe that institutionally store archaeological datasets, like INRAP (Institut National de Recherches Archéologiques Préventives) in France; KNAW-DANS (Data Archiving and Networked Services of the Koninklijke Nederlandse Academie van Wetenschappen) in the Netherlands; and many more. Actually, ARIADNEplus has widened its horizon, including among its providers the already mentioned tDAR in the USA; the Argentinian network of archaeological research centres created under the auspices of CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas); IAA (Israel Antiquities Authority); and Nara (Nara National Research Institute for Cultural Properties) in Japan. At present, 41 partners are involved, coming from 22 EU and EFTA (European Free Trade Area) countries, and from UK, USA, Argentina, Israel, and Japan outside Europe. A continuously increasing number of associate partners is joining the initiative, extending the ARIADNE coverage in practice to all of Europe and beyond.

The process of aggregation consists in the collection of dataset metadata from the data providers, their conversion to a common standard and the inclusion in the catalogue. The original datasets are kept and maintained by the owners.

ARIADNE organises such metadata into a catalogue presently containing about 2,000,000 items, which can be accessed and searched

4 https://www.hadw-bw.de/en/research/research-center/roceeh/home.

5 https://opencontext.org/.

e-ISSN 2724-3923

via the project portal.[6] Search parameters may be defined according to keyword, time, place and data type. The search produces a list of datasets fulfilling such parameters, each one with a short description based on the related metadata stored in the catalogue. Each list item is linked to the original dataset stored at the dataset owner, which can be directly accessed by the user according to the access rules established by the data owner. The catalogue is being continuously updated and new items are added as soon as they become available.

Creating the ARIADNE catalogue has required the setup of common controlled multilingual vocabularies based on Getty's AAT (Art and Architecture Thesaurus, Getty s.d.) and the creation of cross-references for named periods, which are location-dependent, as it is well-known that Iron Age, for instance, covers a different time-span in France, England and Ireland.

From the FAIR perspective, ARIADNE is a one-stop access point to archaeological repositories which includes a find functionality. Interoperability is provided by the use of a common ontology, called AO-Cat (ARIADNE Object Catalogue), which is a subset of the current standard ontology for the domain, CIDOC CRM (Conceptual Reference Model of the Comité International pour la Documentation – International Committee for Documentation of ICOM, the International Council of Museums), usually referred to simply as 'the CRM' (Doerr 2003a; Doerr, Kritsotaki, Boutsika 2011; Doerr, Smith-Ore, Stead 2007). In ARIADNEplus, all contributors' metadata schemas are mapped to > AO-CAT- using tools provided by ARIADNEplus.

As regards re-use, a number of services to re-process the data are being made available to users. They will be operational in a VRE (Virtual Research Environment), i.e. a virtual space within the ARIADNEplus infrastructure where data can be stored, processed and analysed by users. An aspect still under investigation is data reliability, a key factor for re-use. ARIADNE and ARIADNEplus have so far addressed this issue by accepting only data from highly-reputed institutions: but extending data aggregation to new repositories will require investigating how to evaluate the trustworthiness of the data and of their producer.

4 The Semantic ARIADNE Infrastructure

As already mentioned, datasets in ARIADNE are organised according to a general ontology, AO-Cat, a subset of the CRM ontology. AO-Cat includes a limited number of classes that are common to any dataset. AO-Cat is fully documented on the ARIADNE web site.

6 https://portal.ariadne-infrastructure.eu.

The general nature of the AO-Cat structure is determined by the extreme diversity of archaeological datasets. As regards formats, there are texts (usually PDF), images, maps, drawings, tables (e.g. Excel ones) and databases created with different DBMS (Data Base Management Systems). Also their content is extremely diverse: there are excavation reports, sites and monuments descriptions, lists of finds, results of scientific analyses and more. ARIADNE has listed 14 groupings of such subdomains, ranging from a-DNA (ancient DNA) Analyses to Standing Structures. For each sub-domain, an Application Profile is envisaged, i.e. a specification of the AO-Cat ontology. For example, the Application Profile for analytical investigations includes specific classes to better describe the data, such as, among others, Analysis, which describes the kind of analysis used in the investigation, and Sample, which describes the sample being analysed. All the classes and properties used in the ARIADNE Application Profiles are taken from the overall CRM ontology or from one of its extensions[7] such as CRMarcheo, the CRM extension for archaeological excavations; CRMsci, the CRM extension for scientific investigations in general; and CRMdig, the CRM extension for digital objects and activities.

The use of Application Profiles will allow a LOD (Linked Open Data) approach. But it also shows that there is a tension between the factuality implied by the CRM and the semantic representation of abstract concepts and inference processes (Doerr, Kritsotaki, Boutsika 2011; Lippi, Torroni 2016) used in the archaeological discourse. Argumentation and inference appear to be the current frontier of any data-centric archaeological semantic methodology, together with documenting quality, uncertainty and imprecision.

5 Beyond the Aggregation of Archaeological Datasets

Aggregating archaeological datasets from sparse repositories as done in ARIADNE or tDAR is a significant step forward to use and re-use archaeological data. Before they were created, a researcher needed to access many different repositories, some of which did not even provide a search engine. Google search gave no support because it produced too many hits and was unable to go beyond the surface of repositories, without reaching and indexing the datasets they contained.

Nevertheless, further progress is still desirable.

All integrating initiatives rely on metadata that are usually not rich enough to satisfy Kintigh's conditions mentioned above. Kintigh

7 All the CRM documentation is available on the web site: `http://www.cidoc-crm.org/`.

e-ISSN 2724-3923

(2015) shows with an example that knowledge extraction from archaeological texts requires not just recognition of nested relationships but also substantial reasoning to properly assess the information, as well as the ability to analyse texts written in languages different from English: an improvement in the semantics is mandatory. Text mining techniques have been applied to archaeological data in various ways to enrich the metadata originally provided with the records (Richards, Tudhope, Vlachidis 2015; Felicetti et al. 2018), and there are promising efforts to identify nested connections and argumentation built on them (Meghini, Bartalesi, Metilli 2017; Meghini et al. 2018).

An additional complication derives from the intrinsic uncertainty that accompanies all archaeological data. Studies on how to identify and address the fuzziness of archaeological knowledge were initiated almost 20 years ago[8] starting from burial databases and lithics typology, and progressively extending to concepts such as time, place, archaeological site and archaeological 'culture' (Niccolucci, Hermon 2015; Niccolucci, Hermon 2017; Hermon, Niccolucci 2017). Archaeological reasoning and argumentation are likely affected by such fuzziness in an even greater way. Also theoretical considerations as those introduced by Niccolucci, Hermon and Doerr (2015) must be taken into account.

6 Technology: A Quick Overview of the State of the Art and the Need for Innovation

6.1 Semantics

The debate on archaeological argumentation, in the broader domain of archaeological theory, has seen many contributions in the last two centuries, with a peak in the late 20th century. While we will not enter in such discussion, it is clear that the first step to re-use and build on data is documenting argumentation with a neutral approach, keeping into account that the semantic structure of reasoning differs according to the theoretical school authors belong to. Assessment must be left to the user, supported by appropriate documentation. The CRM does not consider such aspects, it deals with actual information stored by researchers, museums, libraries and archives. Thus, it needs to be extended to take account also of how such data were produced, collected, analysed and synthesised. Hints in this direction may come from current studies on narratives, so far semantical-

8 See Niccolucci, D'Andrea, Crescioli 2001; Hermon, Niccolucci 2002; Niccolucci, Hermon 2003; Hermon, Niccolucci 2003; Niccolucci, Hermon 2010.

ly analysed from a literary perspective.[9] Since archaeological reports tell a story about the past, an analysis of narration may give significant insights. Archaeology as a discipline has its reasoning structure that needs to be interpreted and formalised into a model allowing the description of archaeological statements ('interpretation') in formal terms. 'Narrative' modelling has not been applied in archaeology so far. In parallel with such semantic backbone, tools to analyse stored data are equally required.

Such a new ontology will also need to take into account the representation of uncertainty due to the fuzziness of archaeological statements (Niccolucci, Hermon 2017). The application of fuzzy logic in archaeology is motivated by the need to consider the intrinsic imprecision of archaeological concepts. Fuzziness involves time, space and basic concepts such as type, site and culture. It arises when analysing argumentation and the trustworthiness of conclusions. In semantics, fuzziness is introduced with the concept of 'fuzzy ontology' (Cross 2018; Cross, Chen 2018; Di Noia et al. 2019), used in various domains, from medicine (Parry 2004) to news articles (Chang-Shing, Zhi-Wei, Lin-Kai 2005), but not yet applied to the archaeological discourse.

6.2 Machine Learning, Text Mining and Pattern Recognition

Machine learning has been used in the creation/adaptation of tools to search archaeological datasets and to enrich their metadata.

As regards **Text Mining using NLP** (Natural Language Processing), successful examples of initial application in archaeology[10] have demonstrated that they strongly depend on the underlying semantic structure and on multilingual vocabularies. Machine learning is promising great opportunities, especially in identifying arguments and content beyond, or across, different styles, languages, contexts and purpose.

Another field where computer may support archaeological synthesis is **Pattern Recognition** in 2D images. Here the step forward requires going beyond the mere appearance and graphical resemblance, looking instead for stylistic similarity as defined by archaeologists (Bolettieri et al. 2015; Amato, Falchi, Vadicamo 2016). Another topic with a great potential is **3D Shape Recognition**, i.e. the classification of artefacts based on their shape. There is a large number of studies on this subject, those more relevant for archaeology are probably the ones by Tal (2014), Canul Ku et al. (2018) and

9 Ciotti 2016; Meghini, Bartalesi, Metilli 2017; Meghini et al. 2018; Bartalesi, Meghini, Metilli 2017; Bartalesi et al. 2019.

10 Richards, Tudhope, Vlachidis 2015; Felicetti et al. 2018; Esuli, Moreo, Sebastiani 2019.

Hermon et al. (2018), and the research carried out in the EU-funded project GRAVITATE[11] coordinated in 2015-2018 by Michela Spagnolo of CNR (Consiglio Nazionale delle Ricerche) as well as in the ARCHAIDE[12] EU-funded project.

6.3 Virtual Research Environments

A **VRE** (Virtual Research Environment) is a computer framework that virtually mimics a research laboratory, making available in the same virtual space the data and the tools to process them, by individuals or by teams collaboratively working on the same topic (Jeffery et al. 2017). An example of such environment for archaeological research is the one called D4Science (Candela et al. 2014) at ISTI (Istituto di Scienze e Tecnologie dell'Informazione) of CNR in Pisa, Italy. This VRE is being activated for the ARIADNEplus project and will host the services provided by the project to process archaeological data, like the storage of interim results and the reference to their background data, annotation and workflow organisation, text mining, and more.

6.4 Addressing Archaeological Grand Challenges

In conclusion, there are several technologies that may help in managing the archaeological data deluge by organising, synthesising and interpreting them, but they still need an overall methodological organisation to define a data-driven approach to the archaeological research methodology. This is an indispensable step, resulting from the awareness that addressing archaeological Grand Challenges – as defined by Kintigh et al. 2014 – needs to synthesise a huge amount of fragmented information resulting from the convergence of investigations based on many diverse sources and methodologies, such as field campaigns, stylistic analyses, scientific analyses, historical sources and anthropological approaches. As reported in the above-mentioned paper, this concept of archaeological Grand Challenges came after a survey asking the members of the European Association of Archaeologists and the Society of American Archaeologists to indicate the archaeological problems of broad scientific and social interest that could drive cutting-edge research in archaeology for the next decade and beyond. The most compelling and important scientific questions in archaeological research, the Grand Challenges, were then identi-

11 A list of related publications is available here: `http://gravitate-project.eu/?q=content/articles`.

12 `http://www.archaide.eu/`.

fied, elaborating on the answers received and eventually compiling a list of 25 of them. Each one has global significance, requires decisive support from data and involves multidisciplinary collaboration to be solved. The list includes questions about community dynamics, transformation of societies, human-environment interactions and movement and mobility of people, including migrations.

At present, the availability of a wide online access seems reasonably at hand; detailing the contextual information may instead still require substantial research work.

As already mentioned, Kintigh (2015) also shows with an example how misleading a naive approach based on simple word search could be.

Besides the ability to analyse texts written in different languages, a radical change in the semantic approach is therefore mandatory to discover data relevant for re-use which may be buried under several information layers, as they were considered of minor importance for the original research question. Kintigh concludes (2015, 97) that the tools currently available do not support a deep analysis of texts – as published papers or grey literature – which remain one of the most important sources of knowledge:

> Enormous quantities of archaeological information and knowledge are embedded in often-lengthy reports and journal articles […] A number of factors conspire to frustrate synthetic research. They include the problems of discovery and access to archaeological data, the difficulty of integrating data from diverse sources, and the problem of extracting usable data, information, and knowledge from text. [It is necessary] to translate knowledge written in natural language into a state-of-the-art knowledge representation language [that] can be queried by machine reasoning based on formalized basic principles of archaeology. (2015, 97)

Raw scientific data are no less difficult to integrate. For example, Sr isotope analysis is widely used when studying migrations, to identify foreigners – potential immigrants – buried in cemeteries. However, such data might need to be combined with studies on pottery or metalwork, to discover a stop-over of these migrants before they reached their final destination. Viking settlers in Britain may have originated in Denmark or Norway, but it appears from their dress accessories that many made a stopover in Carolingian France or Ireland before reaching England. Migration is frequently a complex phenomenon, not a question of single start- and end-points. Migrations are a good example of how an archaeological Grand Challenge from the above-mentioned list needs a powerful support from data.

In conclusion, data-driven archaeology is not a job for 'parachuted' technologists – here the data science and knowledge organisa-

tion experts – who pop up at some point in time and teach archaeologist how they should organise their methods and way of thinking. It is neither a do-it-yourself machinery, as this would possibly turn into *the blind leading the blinds* of biblical memory (Matt. 15:14). Instead, a *bicycle for two* approach[13] is required: data science for archaeology must be tailored according to the nature of the discipline, the use researchers need to make of them, and how the archaeological discourse is structured.

7 Archaeological Big Data

Archaeological data are increasingly available in digital format but, according to Hugget (2015, 2019), they are messy and complicated by their partial, fragmentary, interpretative nature. Hence, sometimes existing data may not be re-used and incorporated in archaeological interpretation just because they are not identified as relevant or because they are disregarded in the flood of available information. Applications of information technology and proper data organisation endeavour to reduce this risk. The large amount of available data has suggested to address the question in a 'Big Data' framework. As it is well known, the so called 'Big Data' approach collects, stores and analyses very large sets of data, too large to be processed with the usual data processing software. However, the term 'Big Data' applies to archaeological data with a different nuance than it usually has, and their intrinsic diversity may lead to unforeseen results, as argued by Hugget. Studies have demonstrated that inconsistent results may be produced by applying deep learning and AI (Artificial Intelligence) techniques in an irreflexive way (Woodall et al. 2014; Succi, Coveney 2019), replacing causation with correlation, thinking that the numbers speak for themselves and that research can advance even without coherent models or unified theories.[14] In these papers it is also argued that the impact of poor-quality data can increase rather than reduce, as dataset size increases.

In archaeology, 'big' refers more to variety and diversity than to quantity. Archaeology does not create an immense – but conceptually shallow – ocean of data, continuously and rapidly increasing in number, as for example the data used for behaviour analytics on the

13 This definition was originally used by Pollard and Bray for archaeological sciences (2007).

14 This sentence is the Author's synthesis of the conclusions made in Succi's and Coveney's paper (2019).

web or those created by the Internet-of-Things.[15] Such big data can indeed be addressed using powerful computing power and relying on a pretty simple knowledge organisation system. On the contrary, archaeology requires the definition of a complex semantic organisation able to capture and organise the inner meaning of statements, arguments and interpretation. Therefore, data science must adapt to the specific needs of the discipline and focus on refined semantics rather than on large-scale processing only.

There is another common pitfall: the belief that if tools work properly when applied individually, they do the same when used in cascade. It is actually the opposite. A fictitious example will clarify this statement.

Let us assume we have an excellent OCR (Optical Character Recognition) system able to recognise characters from written texts, even handwritten or poorly printed ones, in our case for example inscriptions or historical accounts. Such a system has a success rate of 90%, i.e. it understands correctly 9 characters over 10 and puts them in sequence to form words. Also, let us have a very good text mining system. It can extract, among others, monument descriptions from texts, based on controlled dictionaries, ontologies, and all the required semantic paraphernalia. Only 10% of its extractions turns to be wrong, i.e. it also has a success rate of 90%. Finally, let us assume that a large catalogue of shapes is available, so that the shape of objects like capitals, columns, architraves, pediments and so on are available in the many possible aspects they may have, so that searching e.g. for "column with a Corinthian capital and a noticeable entasis" produces the right picture. Due to some possible ambiguity in the search description, the result is not always as expected, but the system gives the right result in 90% of the cases.

Now, let us create a pipeline formed by the three tools in sequence, so that the outcome of the first one feeds into the second one and this produces a result that is processed by the third one: namely, one inputs the text of some ancient source into such pipeline and gets the picture of the object as the outcome. A naïve attitude would expect that 90% of the results produced by this assemblage are good, but it is not so: error multiplies, so the expected quality of the final result is only $0.9 \times 0.9 \times 0.9 = 0.729$ or 72.9%, i.e. about 30% of the outcomes may be wrong. In other words, the more complicated a process is, and the more passages are involved, the less reliable is the result, unless intermediate results are assessed and cleaned at every step: but this is something our fully automatic mechanism was designed to avoid.

15 This term usually indicates the automatic creation of data by sensors connected to the Internet.

e-ISSN 2724-3923

In conclusion, there is still a long way to go before Big Data techniques, very fashionable today, may apply straightforwardly in archaeological research. This consideration does not imply refusal of such advanced tools, but just critical consideration and the avoidance of an overoptimistic and irreflexive acceptance, based on an acritical approach to technology.

8 Conclusions and Further Work

Although much has been done in the twenty years since Gardin's statement mentioned in the introduction, there is still work to do to achieve an operational data-centric approach in the archaeological research methodology. Accumulating, storing and making openly available archaeological data is a great progress compared to not so many years ago. It saves results for the future, avoids ignoring previous work or re-doing it, and creates an eco-environment of collaborative research. However, without further work it risks making the data deluge more suffocating. If the goal is well summarised by the FAIR principles mentioned above, this cute acronym still hides many unresolved issues.

Access to existing data must be as open as possible while remaining as closed as (strictly) necessary. In archaeology, this is achieved for data coming from research, where openness can be easily enforced by leveraging on funding as most funding agencies do nowadays, requiring the publication of results with open access. The same must be required also from those resulting from administrative activities such as emergency excavations. Once personal and security information is protected, such data must be disclosed to the research community. Intellectual property limitations must not apply to administrative acts, as the reports resulting from emergency excavations, or in general after a reasonable and short embargo period from archaeological discoveries, as it happened in the past by researchers and officers keeping finds hidden for years as they were 'under study'.

Finding data is reasonably suitable with the search system implemented so far, but it needs a substantial improvement as regards the inner connections and argumentation and an in-depth analysis of reports as discussed above. Otherwise, searches will report too many results to be manageable. In other cases, they will still ignore valuable data filtered away by poor metadata, not rich enough to enable the discovery of deeply hidden information as in Kintigh's example mentioned above. Searching must be able to explore the connection of concepts and not just their presence.

Interoperability is probably the FAIR principle where results are most advanced. Possible different perspectives and a vibrant debate do not challenge the global consensus on a shared ontology, with a

handful of exceptions – over which we will draw a veil – diverting from the mainstream and renouncing to global interoperability for vested interests, not worth consideration.

Finally, **Re-use** still requires much work. Quality assessment is a primary concern and a machine-actionable chain of trust is required. While for the other principles the roadmap is clear, for this one exploration is still necessary.

Last but not least, global awareness in the research community is an achievement to be heartily acknowledged. The EAA is undertaking an initiative on these issues and a joint European and US research team has recently proposed an initiative to foster and investigate archaeological data FAIRness. Unfortunately, academic reward for basic work supporting these aspects is still lacking, and for progressing in the career a monograph on some obscure ceramics is still preferred to any global instrument supporting FAIRness as the publication of a digital corpus or the creation of a virtual reference collection.

Bibliography

Aloia, N. et al. (2017). "Enabling European Archaeological Research: The ARIADNE E-Infrastructure". *Internet Archaeology*, 43.

Amato, G.; Falchi, F.; Vadicamo, L. (2016). "Visual Recognition of Ancient Inscriptions Using Convolutional Neural Network and Fisher Vector". *Journal on Computing and Cultural Heritage (JOCCH)*, 9(4), art. 21, 1-24. http://doi.org/10.1145/2964911.

Barcelo, J.A.; Bogdanovic, I. (eds) (2015). *Mathematics in Archaeology*. Boca Raton: CRC Press.

Bartalesi, V.; Meghini, C.; Metilli, D. (2017). "A Conceptualisation of Narratives and Its Expression in the CRM". *International Journal of Metadata, Semantics and Ontologies*, 12(1), 35-46.

Bartalesi, V. et al. (2019). "Introducing Narratives in Europeana: A Case Study" *International Journal of Applied Mathematics and Computer Science*, 29(1). http://doi.org/10.2478/amcs-2019-0001.

Bolettieri, P. et al. (2015). "Searching the EAGLE Epigraphic Material Through Image Recognition via a Mobile Device". Amato, G. et al. (eds), *Similarity Search and Applications. SISAP 2015*. Lecture Notes in Computer Science 9371. Cham: Springer, 351-4. http://doi.org/10.1007/978-3-319-25087-8.

Buard, P.-Y. et al. (2019). "The Archaeological Excavation Report of Rigny: An Example of an Interoperable Logicist Publication". *CIDOC 2018*, September 2018, Heraklion, Greece. https://hal.archives-ouvertes.fr/hal-01892412.

Candela, L. et al. (2014). "Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience". *Proceedings International Symposium on Grids and Clouds (ISGC) 2014, Academia Sinica, Taiwan*. Proceedings of Science PoS(ISGC2014)022. https://doi.org/10.13140/2.1.5035.2327.

Canul Ku, M. et al. (2018). "Classification of 3D Archaeological Objects Using Multi-View Curvature Structure Signatures". *IEEE Access*, December 2018. http://doi.org/10.1109/ACCESS.2018.2886791.

Chang-Shing, L.; Zhi-Wei, J.; Lin-Kai, H. (2005). "A Fuzzy Ontology and Its Application to News Summarization". *IEEE Transactions on Cybernetics*, 35(5), 859-80. https://doi.org/10.1109/TSMCB.2005.845032.

Ciotti, F. (2016). "Toward a Formal Ontology for Narrative". *MATLIT*, 4(1) 29-44. https://doi.org/10.14195/2182-8830_4-1_2.

Cross, V. (2018). "Fuzzy Ontologies: The State of the Art". Gibbs, H.M. (ed.) *Proceedings 2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, IEEE Xplore, 1-8.

Cross, V.; Chen, S. (2018). "Fuzzy Ontologies: State of the Art Revisited". Barreto, G.A.; Coelho, A.L.V. (eds), *Fuzzy Information Processing. NAFIPS 2018. Communications in Computer and Information Science*, vol. 831. Cham: Springer. http://doi.org/10.1007/978-3-319-95312-0_20.

D'Andrea, A.; Niccolucci, F. (2000). "L'archeologia computazionale in Italia: orientamenti, metodi e prospettive". *Archeologia e Calcolatori*, 11, 13-31.

Di Noia, T. et al. (2019). "A Fuzzy Ontology-Based Approach for Tool-supported Decision Making in Architectural Design". *Knowledge and Information System*, 58(1), 83-112. https://doi.org/10.1007/s10115-018-1182-1.

Doerr, M. (2003a). "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata". *AI Magazine*, 24(3), 75.

Doerr, M. (2003b). "The CIDOC Conceptual Reference Model: A New Standard for Knowledge Sharing". *ACM Proceedings Er '07 Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modelling*, vol. 83, 51-6.

Doerr, M.; Kritsotaki, A.; Boutsika, K. (2011). "Factual argumentation – A Core Model for Assertions Making". *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 3(3), Article 8, 1-34. http://doi.org/10.1145/1921614.1921615.

Esuli, A.; Moreo, A.; Sebastiani, F. (2019). "Funnelling: A New Ensemble Method for Heterogeneous Transfer Learning and Its Application to Cross-Lingual Text Classification". *ACM Transactions on Information Systems*, 37, 3, Article 37. https://doi.org/10.1145/3326065.

Felicetti, A. et al. (2018). "NLP Tools for Knowledge Extraction from Italian Archaeological Free Text". Addison, A.C.; Thwaites, H.H. (eds), *Proceedings of the 3rd Digital Heritage International Congress (DigitalHERITAGE) Held Jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*. San Francisco, 2018, 1-8. http://doi.org/10.1109/DigitalHeritage.2018.8810001

Gardin, J.-C. (1980). *Archaeological Constructs*. Cambridge: Cambridge University Press.

Gardin, J.-C. (1999). "Calcul et narrativité dans les publications archéologiques". *Archeologia e Calcolatori*, 10, 63-78.

Gardin, J.-C.; Roux, V. (2004). "The ARKEOTEK Project: A European Network of Knowledge Bases in the Archaeology of Techniques". *Archeologia e Calcolatori*, 15, 25-40.

Getty (s.d.). *Art & Architecture Thesaurus*. https://www.getty.edu/research/tools/vocabularies/aat/.

Hermon, S. et al. (2018). "An Integrated 3D Shape Analysis and Scientific Visualization Approach to the Study of a Late Bronze Age Unique Stone Object

from Pyla-Kokkinokremos, Cyprus". *Digital Applications in Archaeology and Cultural Heritage*, vol. 10, Article e00075. https://doi.org/10.1016/j.daach.2018.e00075.

Hermon, S.; Niccolucci, F. (2002). "Estimating Subjectivity of Typologists and Typological Classification with Fuzzy Logic". *Archeologia e Calcolatori*, 13, 217-32.

Hermon, S.; Niccolucci, F. (2003). "A Fuzzy Logic Approach to Typology in Archaeological Research". Doerr, M.; Sarris, A. (eds), *The Digital Heritage of Archaeology.* Athens: Archive of Monuments and Publications, Hellenic Ministry of Culture, 307-10.

Hermon, S.; Niccolucci, F. (2017). "Formally Defining the Time-space Archaeological Culture Relation: Problems and Prospects". *Archeologia e Calcolatori*, 28, 93-108.

Huggett, J. (2015). "A Manifesto for an Introspective Digital Archaeology". *Open Archaeology*, 1(1), 86-95. https://doi.org/10.1515/opar-2015-0002.

Huggett, J. (2019). "Delving into Data Reuse". https://introspectivedigitalarchaeology.com/2019/10/10/delving-into-data-reuse.

Jeffery, K.G. et al. (2017). "A Reference Architecture for Virtual Research Environments". Gäde, M.; Trkulja, V.; Petras, V. (eds), *Everything Changes, Everything Stays the Same? Understanding Information Spaces = Proceedings of the 15th International Symposium of Information Science (ISI 2017)* (Berlin, Germany, March 13-15, 2017). Glückstadt: Verlag Werner Hülsbusch, 76-88.

Kenny, A. (1991). "Technology and Humanities Research". Katzen, M. (ed.), *Scholarship and Technology in the Humanities*. London: Bowker Saur, 1-10.

Kintigh, K.W. (2006). "The Promise and Challenge of Archaeological Data Integration". *American Antiquity*, 71(3), 567-78. https://doi.org/10.2307/40035365.

Kintigh, K.W. (2015). "Extracting Information from Archaeological Texts". *Open Archaeology*, 1, 96-101. https://doi.org/10.1515/opar-2015-0004.

Kintigh, K.W. et al. (2014). "Grand Challenges for Archaeology". *American Antiquity*, 79(1), 5-24. https://doi.org/10.7183/0002-7316.79.1.5.

Lippi, M.; Torroni, P. (2016). "Argumentation Mining: State of the Art and Emerging Trends". *ACM Transactions on Internet Technology (TOIT)*, 16(2), 1-25. https://doi.org/10.1145/2850417.

Marlet, O. et al. (2019). "The Archaeological Excavation Report of Rigny: An Example of an Interoperable Logicist Publication". *Heritage* 2019, 2(1), 761-73. https://doi.org/10.3390/heritage2010049.

McManamon, F.P.; Kintigh, K.W.; Brin, A. (2010). "Digital Antiquity and the Digital Archaeological Record (tDAR): Broadening Access and Ensuring Long-Term Preservation for Digital Archaeological Data". *The CSA Newsletter*, 23(2).

Meghini, C.; Bartalesi, V.; Metilli, D. (2017). "Using Formal Narratives in Digital Libraries". Grana, C.; Baraldi, L. (eds), *Digital Libraries and Archives = Proceedings of the 13th Italian Research Conference on Digital Libraries, IRCDL 2017* (Modena, Italy, January 26-27, 2017). Revised Selected Papers. Communications in Computer and Information Science 733. Cham: Springer, 83-94. https://doi.org/10.1007/978-3-319-68130-6_7.

Meghini, C. et al. (2017). "ARIADNE: A Research Infrastructure for Archaeology". De Santo, M.; Niccolucci, F.; Richards, J.D. (eds), *Journal on Computing and Cultural Heritage, Special Issue on Research Infrastructures*, 10(3), August, Article No. 18, 1-27. http://doi.org/10.1145/3064527.

Meghini, C. et al. (2018). "A Software Architecture for Narratives". Serra, G.; Tasso, C. (eds), *Digital Libraries and Multimedia Archives = Proceedings of*

the 14th Italian Research Conference on Digital Libraries, IRCDL 2018 (Udine, Italy). Communications in Computer and Information Science 806. Cham: Springer, 23-29. https://doi.org/10.1007/978-3-319-73165-0_3.

Niccolucci, F. (2018). "Integrating the Digital Dimension into Archaeological Research: The ARIADNE Project". *Post-Classical Archaeologies*, 8, 281-7.

Niccolucci, F.; D'Andrea, A.; Crescioli, M. (2001). "Archaeological Applications of Fuzzy Databases". Stančič, Z.; Veljanovski, T. (eds), *Computing Archaeology for Understanding the Past*. Oxford: Archaeopress, 107-16. BAR International Series 931.

Niccolucci, F.; Hermon, S. (2003). "La logica fuzzy e le sue applicazioni alla ricerca archeologica". *Archeologia e Calcolatori*, 14, 97-110.

Niccolucci, F.; Hermon, S. (2010). "A Fuzzy Logic Approach to Reliability in Archaeological Virtual Reconstruction". Niccolucci, F.; Hermon, S. (eds), *Beyond the Artefact. Digital Interpretation of the Past*. Budapest: Archaeolingua, 28-35.

Niccolucci, F.; Hermon, S. (2015). "Time, Chronology and Classification". Barcelo, Bogdanovic 2015, 257-71.

Niccolucci, F.; Hermon, S. (2016). "Representing Gazetteers and Period Thesauri in Four-dimensional Space-Time". *International Journal on Digital Libraries*, 17(1), 63-9. https://doi.org/10.1007/s00799-015-0159-x.

Niccolucci, F.; Hermon, S. (2017). "Expressing Reliability with CIDOC CRM". *International Journal on Digital Libraries*, 18(4), 281-7. https://doi.org/10.1007/s00799-016-0195-1.

Niccolucci, F.; Hermon, S.; Doerr, M. (2015). "The Formal Logical Foundations of Archaeological Ontologies". Barcelo, Bogdanovic 2015, 86-99.

Niccolucci, F.; Richards, J.D. (2013). "ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe". *International Journal of Humanities and Arts Computing*, 7(1-2), 70-88. https://doi.org/10.3366/ijhac.2013.0082.

Niccolucci, F.; Richards, J.D. (2019). "Ariadne and ARIADNEplus". Richards, Niccolucci 2019, 7-26.

Nuninger, L. at al. (2020). "Linking Theories, Past Practices, and Archaeological Remains of Movement through Ontological Reasoning". *Information* 2020, 11(6), 338. http://doi.org/10.3390/info11060338.

Parry, D. (2004). "A Fuzzy Ontology for Medical Document Retrieval". Hogan, J. et al. (eds), *ACSW Frontiers '04: Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*. Darlinghurst: Australian Computer Society, 121-6.

Pollard, A.M.; Bray, P. (2007). "A Bicycle Made for Two? The Integration of Scientific Techniques into Archaeological Interpretation". *Annual Review of Anthropology*, 36(2007), 245-59. https://doi.org/10.1146/annurev.anthro.36.081406.094354.

Richards, J.D. (1997). "Preservation and Re-use of Digital Data: the Role of the Archaeology Data Service". *Antiquity*, 71(274), 1057-9. https://doi.org/10.1017/S0003598X00086014.

Richards, J.D. (2017). "Twenty Years Preserving Data: A View from the United Kingdom". *Advances in Archaeological Practice*, 5(3), 227-37. https://doi.org/10.1017/aap.2017.11.

Richards, J.D.; Niccolucci, F. (eds) (2019). *The ARIADNE Impact*. Budapest: Archaeolingua.

Richards, J.D.; Tudhope, D.; Vlachidis, A. (2015). "Text Mining in Archaeology: Extracting Information form Archaeological Records". Barcelo, Bogdanovic 2015, 240-54.

Succi, S.; Coveney, P.V. (2019). "Big Data: The End of the Scientific Method?". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2142). http://doi.org/10.1098/rsta.2018.0145.

Tal, A. (2014). "3D Shape Analysis for Archaeology". Ioannides, M.; Quak, E. (eds), *3D Research Challenges in Cultural Heritage*. Lecture Notes in Computer Science 8355. Berlin; Heidelberg: Springer, 50-63. https://doi.org/10.1007/978-3-662-44630-0_4.

Wilkinson, M.D. et al. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data*, 3 art. http://doi.org/160018.10.1038/sdata.2016.18.

Wilkinson, M.D. et al. (2019). "The Addendum: FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data*, 6 art. 6. http://doi.org/10.1038/s41597-019-0009-6.

Woodall, P. et al. (2014). "An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics". MIT Information Program (ed.), *Proceedings of the 19th International Conference on Information Quality (ICIQ 2014): Big Data: Management and Data Quality, Xi'an (China)*. Red Hook (NY): Curran Associates, 24-33.

Zadora-Rio, E. et al. (2020). "L'église de Rigny et ses abords. De la *colonia* de Saint-Martin de Tours au transfert du centre paroissial (600-1865)". https://www.unicaen.fr/puc/rigny/accueil.