pca

european journal of
postclassicalarchaeologies

volume 8/2018

SAP Società Archeologica s.r.l.

Mantova 2018

Post-Classical Archaeologies (PCA) is an independent, international, peer-reviewed journal devoted to the communication of post-classical research. PCA publishes a variety of manuscript types, including original research, discussions and review articles. Topics of interest include all subjects that relate to the science and practice of archaeology, particularly multidisciplinary research which use specialist methodologies, such as zooarchaeology, paleobotany, archaeometallurgy, archaeometry, spatial analysis, as well as other experimental methodologies applied to the archaeology of post-classical Europe.

Submission of a manuscript implies that the work has not been published before, that it is not under consideration for publication elsewhere and that it has been approved by all co-authors. Each author must clear reproduction rights for any photos or illustration, credited to a third party that he wishes to use (including content found on the Internet). For more information about **ethics** (including plagiarism), copyright practices and guidelines please visit the web site www.postclassical.it.

PCA is published once a year in May, starting in 2011. Manuscripts should be submitted to **editor@postclassical.it** in accordance to the guidelines for contributors in the webpage http://www.postclassical.it

Post-Classical Archaeologies's manuscript **review process** is rigorous and is intended to identify the strengths and weaknesses in each submitted manuscript, to determine which manuscripts are suitable for publication, and to work with the authors to improve their manuscript prior to publication.

This journal has the option to publish in **open access**. For information please visit the web site www.postclassical.it

How to **quote**: please use "PCA" as abbreviation and "European Journal of Post-Classical Archaeologies" as full title.

Cover image: vectorialised cadastre of Borgo Rudena, Padova (F. Giacomello).

"Post-Classical Archaeologies" is indexed in Scopus. It was approved on 2015-05-13 according to ERIH PLUS criteria for inclusion. Classified A by ANVUR (Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca).

DESIGN
Paolo Vedovetto

PUBLISHER
SAP Società Archeologica s.r.l.
Strada Fienili 39/a, 46020 Quingentole, Mantova
www.archeologica.it

PRINTED BY
Tecnografica Rossi, Via I maggio, Sandrigo (VI)

Authorised by Mantua court no. 4/2011 of April 8, 2011

For subscription and all other information visit the web site www.postclassical.it

ISSN 2039-7895

pca
european journal of
postclassicalarchaeologies
volume 8/2017

CONTENTS PAGES

BEYOND THE THEME

PROJECT

REVIEWS

PCA 8 (2018) ISSN: 2039-7895 (pp. 325-336)
Post - Classical Archaeologies

project

Integrating the digital dimension into archaeological research: the ARIADNE project

FRANCO NICCOLUCCI

VAST-LAB, PIN - University of Florence, University Pole "Città di Prato", P.zza G. Ciardi, 25, 59100 Prato,
franco.niccolucci@gmail.com

The ARIADNE project, started in 2013, has created a catalogue of archaeological datasets in Europe registering about 2.000.000 items including reports, images, drawings, GIS and databases. The search functionalities it provides enable users to select such datasets according to criteria based on time, location and content, and then to access source data. Since its completion in 2016 it has become popular among archaeologists. An improved version is currently in the plans. It will create cloud-based virtual research environments where researchers can collaborate and analyse data to re-use them in their investigations.
Keywords: digital archaeology, archaeological datasets, research infrastructures, virtual research environments, ARIADNE registry

Il progetto ARIADNE, iniziato nel 2013, ha creato un catalogo dei dataset *archeologici in Europa registrando circa 2.000.000 archivi che comprendono rapporti, immagini, disegni, GIS e* database*. Le funzionalità di ricerca che mette a disposizione permettono agli utenti di selezionare tali archivi in base a criteri di ricerca basati su tempo, luogo e contenuto, e quindi di accedere alle fonti originali. Fin dal suo completamento nel 2016 è diventato popolare fra gli archeologi. Una versione migliorata è nei piani: questa creerà ambienti virtuali di ricerca, dove gli utenti potranno collaborare e analizzare i dati per riutilizzarli nelle loro ricerche.*
Parole chiave: *archeologia digitale,* dataset *archeologici, infrastrutture di ricerca, ambienti di ricerca virtuali, registro ARIADNE*

1. Introduction

The ARIADNE project was born in 2012 from the idea that a large part of the results of archaeological investigations was not easily available to researchers, although produced in digital format and hence easily accessible, at least in theory. Apart from the traditional scientific publi-

Received: 03-03-2018 - Accepted: 12-03-2018 - Revised: 20-03-2018

cations on journals and the communications at conferences published in proceedings, an increasing amount of reports, databases, GIS and visual data were dispersed in a myriad of on-line repositories, difficult to find and access. On the other hand, the concept of sharing data and re-using those produced by others was well accepted by the archaeological community (Niccolucci, Richards 2013, p. 74). Data sharing may indeed be considered the 21st century equivalent of what Gerhard about two centuries ago called "the continuous and mutual assistance of many … who mutually share their discoveries and knowledge", stating that without it, in archaeology "the efforts of one person alone may never succeed" (Gerhard 1829).

There were many obstacles to achieve this objective. As already mentioned, archaeological datasets were published on the web as a *corpus* fragmented in many different Internet locations and were organized in the most diverse ways. Modern borders divided what belonged to the same archaeological region, and split accordingly their digital archives, as for example the Western Mediterranean Roman Provinces, corresponding at present to at least three different countries. Language and terminology diversity made Google-style searching ineffective, and not usable to access data stored in databases. Last but not least, the ARENA project had demonstrated that a naïve search based on named time periods, like "Iron Age", would produce incoherent results: this was jokingly described with the example of a ship crossing the Channel in AD 30, starting the trip in the (Gaul) Roman period and ending it in the (Britain) Iron Age, thus traveling not only across the sea, but also backwards in time.

The opportunity of attempting the integration of archaeological data available on-line was offered by an EU-funded program aimed at so-called integrating activities for research infrastructures, i.e. the combination of facilities, human resources and services operating in a specific scientific domain. The EU fostered the aggregation of regional, national or international initiatives that were already providing access to their data, were supported by expert researchers willing to participate in such integration process, and could contribute to the implementation of valuable web services, first of all an advanced search system on the knowledge base resulting from such integration. Such support was provided in any scientific domain, including archaeology, through a competitive call.

Thus a proposal was submitted, supported by 23 partners from 16 European countries. The partnership included some research centers providing the necessary technical knowledge. All the others were national archaeological institutions already managing an archaeological data infrastructure of national importance. Among others, the partnership in-

cluded the Archaeological Data Service at the University of York (UK), a forerunner of archaeological data deposit and sharing; several institutions in charge of managing national archaeological archives such as INRAP in France, DANS in the Netherlands and MIBACT in Italy; and the archaeological institutes of the Academies of Sciences of several Eastern and Central European countries.

The proposal was successful, and the project started on 1st February 2013 with a planned duration of four years.

2. The first stage of the ARIADNE project

The key idea of ARIADNE was that integrating does not mean assembling. The project did not intend to bring all the European archaeological data in the same mammoth repository: it would have been unmanageable, expensive and inefficient. Instead, the project was based on the principle of keeping the data where they were and under the control of their owners, facilitating discovery and access through a centralized catalogue with efficient search functionalities. Thus integration would have been based on a lightweight Registry, with links to the original datasets. Users could find the relevant datasets searching the Registry and filtering what was of their interest, and then they could access the original data at the owner's repository via the links provided. This also solved the difficult question of the permission rules to access archaeological data, which often vary from country to country and from repository from repository: since the last step of data access was under the control of the repository owner, different access regulations could be applied to users in a straightforward way.

The ARIADNE project had to face a number of technical issues, due to the independent and uncoordinated creation of the repositories it aimed to integrate.

The first one addressed was the structure of the Registry, i.e. the definition of what kind of information was necessary to identify so diverse datasets, and sufficient to filter out in searches all unwanted items, but keeping in all the desired ones.

For this purpose the project set up a data model, with a preliminary version used in the beginning and a final, improved one at the end of the project. The current version of the data model includes the essential information for each dataset, grouped in four sections: Actors, Physical Objects, Digital Objects, and Events. The Actors section concerns people and institutions involved, with their roles: archaeologists, research centers, data repository managers, and so on. Physical Objects include

all artifacts documented, from finds to monuments and sites. The Digital Objects section describes datasets involved, their formats and so on. Finally, Events includes all activities related to the Physical Objects and to the Digital Objects, like excavations, scientific analyses, dating, as well as creating the digital documentation, 3D scanning or any other data production, processing or management. Such data model, developed within the PARTHENOS project, can be applied to the most different file formats: text reports, images, drawings, and any combination/collection of the above, such as, for example, a folder containing the data documenting an excavation, i.e. excavation diaries, drawings, photos, lists of finds, and so on. The data model describes the physical and research context as well as the digital operations that produced the data.

Such summary information on datasets or dataset collections is obtained from the metadata already available for them, with no human intervention except when the system evidences a mistake, what happened in a small percentage of the total. Often the error or lack of compliance to the standard model was systematic and could be easily corrected: for example, this was the case of coordinates, with some partners choosing to indicate latitude first while others put the value of longitude before the other one. Once this was detected, corrections could be automatic and straightforward. Nevertheless, a very small number of errors still remain: recently the location pointer of a dataset concerning Aquileia was noticed to appear in the Atlantic Ocean. Such errors are inevitable in large catalogues and derive from a mistake in the original dataset. They

Fig. 1. A screenshot of the ARIADNE portal.

Fig. 2. The result list from the ARIADNE portal.

are corrected in the original data and in the catalogue as soon as they are noticed or reported by users. However, such mistakes only affect searching the catalogue according to the criteria where the mistake is present in the original data, e.g. the location in the previous example, and have no adverse effect in all other searches.

Fig. 1 shows the ARIADNE portal home page. Fig.2 shows a result page; no search parameter was input to show the total number of datasets in the catalogue (top center), so the result list includes all the catalogue items.

The ARIADNE catalogue at present contains almost 2.000.000 items. It is searchable according to several facets: when, where, what are the main ones, plus a keyword search and information about the data owner and the publisher. Search conditions may be independently set for each one of them, using a visual interface for the time span and the location involved, and controlled lists for named periods and subjects.

The search result is a list of short descriptions, which may be further refined as required, applying again additional filters. Each list item consists in the title and concise essential information about the content referred. Such items are linked to the original data stored at the data owner, which can be directly accessed via the link. Thus users may access data only when they seem more relevant to their interest or research question. In sum, the system was defined in a TV interview for the general public as “a Google for archaeology, but an intelligent one”: indeed, it is much more.

3. Implementation details and issues

A number of technical solutions were devised to develop the ARIADNE system.

The main choice concerned the semantic data model, i.e. how to organize the knowledge contained in the data. This requires agreeing on a metadata system able to express the meaning of individual pieces of information, which otherwise would have little or no significance. For example, in the expression "gold cup" we can recognize a material (the gold) and a man-made object (the cup) of a specific type, an object for drinking. Thus it has some similarity with a "ceramic mug" as both objects are used for drinking, although different in shape and material. It also has similarity with "silver ring", both being objects made in precious material. Organizing knowledge requires thus to categorize the main concepts to be used in the archaeological discourse and in the documentation, organized in a hierarchy of entities (a so-called *ontology*) going from the most general ones to more detailed ones.

It was agreed that the main distinction is between *Temporal Entities*, i.e. phenomena that happen during an extent of time as, for example, *Events* (the destruction of Pompeii), *Activities* (the archaeological excavations at Pompeii), and so on; and *Persistent Items*, i.e. items that have a persistent identity during their existence, as a *Person*, a *Thing*, a *Man-Made Object* and so on. Both these categories are then further specialized in more detailed concepts, and linked through properties, stating for example that a *Person* (e.g. an archaeologist) participates in an *Activity* (e.g. an excavation). A complete system of this kind already exists for cultural heritage, the CIDOC Conceptual Reference Model, briefly called the CRM. The CRM is widely accepted and has the characteristic of being extensible, i.e. it can incorporate new concepts as refinements of already existing ones.

Within ARIADNE, a team examined if the CRM includes all the concepts that are necessary and useful to organize the archaeological knowledge and if any of the existing concepts is superfluous or useless in archaeology, possibly introduced in the CRM to be used in other disciplines. For the latter the solution is simple: it is sufficient not to use them for archaeological documentation. For the missing entities, instead, some general concept must be specialized, creating a new entity as a special case of an existing one, for example creating the entity "drinking object" as a special case of man-made ones. This specialization process maintains the compatibility among different metadata systems: for example, one system using the newly created label of "drinking objects" is compatible with another one not using it, since such drinking objects are just a special case of CRM's *Man-Made Object* concept that is used in

both, just expressed with a greater level of detail. Thus the coarser classification of such objects as man-made ones is valid for both and is not altered by the specialization existing in one only. In proposing extensions, it is necessary to resist to the temptation to make too many distinctions and introduce too many detailed concepts, what would prevent recognizing relationships and connections. When to stop in detailing is a matter of experience and efficiency, where the experience of domain experts combines with the expertise of knowledge engineers. In ARIADNE the collaboration of these two different skills matched perfectly, and produced an extension of the core CRM called CRMarcheo, in which the new concepts introduced mainly concern the excavation activity.

The discussion about the best way of organizing the data led also to address more theoretical issues, an activity often considered a futile exercise in the archaeological discourse. This is possibly the case when communication among humans is concerned, but it is unavoidable when using a foolishly precise computer system. For example, such issues concern fundamental categories as time and space. The statement "Verucchio is an iron age site in Emilia-Romagna, Italy" is correct, as the Verucchio territory hosts Villanovian necropolises. But it is impossible to convert it into a precise statement as required in a computer search: does it concern the whole territory of the present Verucchio municipality? Are finds and remains spread in the whole time range from the 12th to the 6th century BC, corresponding to Iron Age in Italy? Moreover, are space and time references independent from each other? The answer is clearly no, as already mentioned about named time periods, which vary according to places, and place names, which correspond to different space extents during time.

The theoretical debate about the best way of describing in a machine-understandable way concepts that may appear clear to humans but are misleading when processed in a computer, produced several iteresting results (Felicetti *et al.* 2015; Ronzino *et al.* 2016; Niccolucci 2016; Niccolucci, Hermon 2016; Hermon, Niccolucci 2017; Niccolucci, Hermon 2017; Niccolucci 2017) and is still open. It often led to reconsider also the logical foundations of such concepts, questioning and discussing issues such as the interrelations between a concept and the terms used to define it, as noticed above for time/space; the intrinsic fuzziness (what is the border of an archaeological site?) of space and time (Niccolucci, Hermon 2015; Niccolucci *et al.* 2015); the distinction between a physical object and its conceptual value; and more. This debate is still in progress.

Thus, the technology behind ARIADNE was one of the cases in which a practical necessity – data organization – generated theoretical questions and brought to reconsider some fundamental categories of the archaeological documentation.

Multilingual thesauri were also required to enable searching in a mass of documents written in different languages. The issue was managed using existing multilingual thesauri as the Getty ones and others available for national languages, which were matched and merged to enable subjects expressed in all the languages used in the datasets, potentially as many as the official EU languages.

For place names, the system used a modern gazetteer, with the awareness that a historical one like Pleaides might perhaps have been preferential. However, most of the place names used in the datasets referred to the modern name of the archaeological sites involved, so we believe that Geonames was more suitable for the kind of data concerned.

Finally, period names were organized using the PeriodO system, which is an approximated matching system between places and the chronology of period names. Although not perfect, its simplified approach gives good results to mix and match named periods. For each "region", i.e. an area in Europe where there is enough chronological uniformity, it uses a table where period names are matched to corresponding time spans, created with the support of experienced archaeologists. Its limits stay in the fact that the areas of such "uniform" regions also depend on time: for example, assigning time-spans to named periods in Italy depends on what is considered as "Italy" through time. Nevertheless, the system substantially improves time-based search criteria.

4. Concluding the Early ARIADNE Period

At the beginning of 2016 the ARIADNE Portal was fully operational. It enabled searching across a large number of digital archaeological repositories with criteria based on time, space and content. Access to the catalogue is – and will remain – free. In the last year of operations, more than 10.000 individual users accessed the system. Feedback from them was very positive: supporting comments were collected from individual users, research institutions, national agencies and archaeological associations, like EAA – the European Association of Archaeologists – and EAC – the European Archaeological Council. A number of institutions not included in the original participant consortium asked to join in to add their repositories to the catalogue, extending the coverage to more European countries. An independent survey (Münster 2017) mentions ARIADNE as the second most important project in digital archaeology.

ARIADNE fully achieved its objectives, both as regards data integration and in creating a community of users. Still much work remains to do. Maintaining the system operational requires limited resources, which

have been volunteered by the original participants. New additions require instead additional work and resources, and are postponed to a second phase of the project, as described below.

The plans for continuation include extending the geographic coverage to countries not present in the initial version; incorporating scientific datasets, like those concerning archaeological sciences and dating; improving the treatment of geographic data across different GIS systems; improving integration; and providing better services for data post-processing. Actually, the latter were present in the original system in an experimental way, but their full implementation requires additional technical resources for development.

6. New avenues for data-driven research

As already mentioned, the original ARIADNE idea was conceived in early 2012. In the meantime technology has progressed and EU strategies had developed the idea of a cloud environment for scientific purposes, called the European Open Science Cloud (EOSC), supporting data-driven science and providing a framework to share data and to re-use them for new investigations. This perspective fits very well with the ARIADNE spirit: the project aims at enabling archaeological research to take full advantage of the opportunities offered by the digital revolution. The EOSC identifies domains in which research strongly builds on digital data, as physics and biology, usually characterized by very large datasets (the so-called Big Data); and others where the support of data to research is currently minimal, typically using small datasets, called in a somehow derogatory way "the long tail of science". As a matter of fact, the "long tail" is a typical feature of the contemporary digital society. In economy, it created successful phenomena such as Amazon, serving a community of customers wishing to buy items for which there is a very small individual demand but altogether create a huge market. *Mutatis mutandis*, the same could happen in digital archaeology: there is a huge number of small datasets, each one probably interesting few specialized researchers, but altogether creating an unprecedented knowledge base.

Besides sharing knowledge, the cloud environment offers two additional features that are of the outmost interest also for archaeologists. Firstly, it enables and fosters collaborative research at distance. The typical 19th and early 20th century scenario of archaeologists as solo players, like explorers of the past, has been gradually but decisively superseded by team work, with various skills and competences involved, and is as far from the archaeological current good practice as is the In-

diana Jones stereotype. Having a place – although a virtual one – where fieldwork notes and reports, laboratory activity, scientific analyses and the results of previous investigations could be stored, shared and jointly analized would substantially improve the efficiency and the collaboration of interdisciplinary teams, both if members are located in the office next door or thousands of kilometers away: this is something that we can experience in our everyday life with smartphone applications, mailing lists and shared online storage. Secondly, this common virtual area could be endowed with research services enabling users to process results and actually do data-driven research. The shared storage would thus become a Virtual Research Environment, to be further combined with access to a digital library for bibliography and to an ARIADNE-like catalog of what others have produced and published from their investigations.

Significant step forwards have been achieved since early 2017, conclusion of the funded project. Developing a tool already present in ARIADNE, a team from the former ARIADNE partnership has developed a Natural Language Processing (NLP) – Named Entity Recognition (NER) tool that operates on archaeological texts. This tool is able to identify in a text, for example an excavation report, the named entities mentioned in it: for example, it can recognize "Pompeii" as a location, "*opus incertum*" as a construction technique, that "samian ware" and "*terra sigillata*" refer to the same type of Roman red pottery, and so on. The results of such NER, based on the data model developed in ARIADNE, create additional metadata, which enrich the machine-readable knowledge about the dataset and allow more refined searches on the whole catalogue. This tool works natively in a cloud environment and has already been developed with the

Fig. 3. The results of the application of TEXTCROWD to an archaeological report.

code-name of TEXTCROWD in the framework of an EOSC-related project called EOSC-pilot, as a demonstrator of how the cloud can be useful for researchers in the archaeological domain. The demonstrator at present works on Italian texts, but extension to other languages for which linguistic rules have been developed, like English, German, Spanish and more, is relatively straightforward. Fig. 3 shows the result of the processing applied to an excavation report; entities identified are evidenced in color only for the sake of display, e.g. green = actors, yellow = sites, cyan = materials, and so on; they are actually stored in machine-readable format.

7. Towards ARIADNEplus: conclusions and further work

Recently a new stage has been planned for ARIADNE. Titled ARIADNEplus, this evolution of the ARIADNE project seeks to achieve substantial improvements as regards integration, searchability and services. The system will be moved to a cloud environment, at present identified with the D4Science cloud managed by CNR-ISTI. This will happen seamlessly: users will not notice any difference with the present data organization. Then datasets metadata will be enriched using an improved NER tool like TEXTCROWD. On the other hand, D4Science incorporates a GIS cloud server to which all the archaeological geographic information will be referred, in order to create a global geographic catalogue. Database structure will be analyzed and mapped on the ARIADNE CRM model. This approach will ultimately create an integrate Linked Open Data system enabling navigation across datasets in addition to improved search functionalities.

Extensions are also planned as regards content. Besides extending the geographic coverage to practically all Europe, scientific datasets will be fully incorporated as well. These include palaeobiology, environmental archeology, archaeometry and more. The new partnership includes now 41 partners and is open to further collaborations. It will also be in charge of supporting the creation of deposit systems for archaeological data where these are not available, then integrating them in the overall catalogue. In the future, perhaps in a third phase of the ARIADNE initiative, this will allow individual research datasets to be added to the catalogue, through a review and accept procedure that is still to be set up. With a 2030 horizon, the digital contribution to archaeological research will become common practice. It will be instrumental in moving archaeological research methods into the 3rd millennium, without abandoning its foundational characteristic of study of the past through the analysis of material culture, but integrating it with the innovative support of digital technology.

References

J.A. Barcelo, I. Bogdanovic (eds) 2015, *Mathematics and Archaeology*, Boca Raton.

A. Felicetti, F. Murano, P. Ronzino, F. Niccolucci 2015, *CIDOC CRM and epigraphy: A hermeneutic challenge*, "CEUR Workshop Proceedings", 1656, pp. 55-68.

F.W.E. Gerhard 1829, *Osservazioni preliminary*, "Annali dell'Instituto di Corrispondenza Archeologica", 1, p. 3.

S. Hermon, F. Niccolucci 2017, *Formally defining the time-space-archaeological culture relation: problems and prospects*, "Archeologia e Calcolatori", 28, pp. 93-108.

S. Münster 2017, *A Survey on Topics, Researchers and Cultures in the Field of Digital Heritage*, "ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences", Volume IV-2/W2, pp. 157-162.

F. Niccolucci 2016, *Documenting archaeological science with CIDOC CRM*, "International Journal on Digital Libraries", 18(3), pp. 223-231.

F. Niccolucci 2017, *Extending, mapping, and focusing the CIDOC CRM. Introduction to the special issue*, "International Journal on Digital Libraries", 18(4), pp. 251-252.

P. Ronzino, A. Felicetti, F. Niccolucci, M. Doerr 2016, *CRMba a CRM extension for the documentation of standing buildings*, "International Journal on Digital Libraries", 17(1), pp. 71-78.

F. Niccolucci, S. Hermon 2015. *Time, chronology and classification*, in Barcelo, Bogdanovic 2015, pp. 257-271.

F. Niccolucci, S. Hermon 2016, *Representing gazetteers and period thesauri in four-dimensional space–time*, "International Journal on Digital Libraries", 17(1), pp. 63-69.

F. Niccolucci, S. Hermon 2017, *Expressing reliability with CIDOC CRM*, "International Journal on Digital Libraries", 18(4), pp. 281-287.

F. Niccolucci, S. Hermon, M. Doerr 2015, *The formal logical foundations of archaeological ontologies*, in Barcelo, Bogdanovic 2015, pp. 86-99.

F. Niccolucci, J. D. Richards 2013, *ARIADNE: Advanced Research Infrastructure for Archaeological Dataset Networking in Europe*, "International Journal of Humanities and Arts Computing", 7(1-2), pp. 70-88.

Websites

ARENA, http://ads.ahds.ac.uk/arena/

CRM, http://www.cidoc-crm.org/

CRMarcheo, http://www.cidoc-crm.org/crmarchaeo/

D4Science, https://www.d4science.org/

EOSC,https://ec.europa.eu/research/openscience cr/pdf/eosc_declaration.pdf

Getty, http://www.getty.edu/research/tools/vocabularies/index.html

PARTHENOS, http://www.parthenos-project.eu/

PeriodO, http://perio.do/

TEXTCROWD,http://eoscpilot.eu/science-demos /textcrowd

All web references visited on 30/4/2018.