White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247028/

Version: Published Version

Article:
Wright, Holly, Gattiglia, Gabriele and Anichini, Francesca (2026) AI and Archaeological Pottery Identification: Reflections on the importance of open and persistent data. Internet Archaeology. ISSN: 1363-5387

https://doi.org/10.11141/ia.74.12

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

This PDF is a simplified version of the original article published in Internet Archaeology under the terms of the Creative Commons Attribution 3.0 (CC BY) Unported licence. Enlarged images, models, visualisations etc which support this publication can be found in the original version online. All links also go to the online original.

Please cite this as: Wright, H., Gattiglia, G. and Anichini, F. 2026 AI and Archaeological Pottery Identification: Reflections on the importance of open and persistent data, Internet Archaeology 74. https://doi.org/10.11141/ia.74.12

AI and Archaeological Pottery Identification: Reflections on the importance of open and persistent data

Holly Wright, Gabriele Gattiglia and Francesca Anichini

This article is a discussion of the work of the Archaeological Automatic Interpretation and Documentation of cEramics (ArchAIDE) project, in the context of the new collaborative work now being undertaken in the AUTOMated enriched digitisation of Archaeological liThics and cerAmics (AUTOMATA) project, including partners from both ArchAIDE and the Arch-I-Scan Project (and described in the other articles within this volume).

Completed in 2019, the purpose of the ArchAIDE project was to develop new technologies and methodologies to optimise and economise the identification process for archaeological pottery. It resulted in a range of outputs, primarily a new proof-of-concept mobile app that streamlined the practice of pottery recognition by type using Artificial Intelligence (AI). ArchAIDE represented one of the earliest contributions to research in the area of automated image recognition of archaeological material, but more importantly, its development was driven entirely by the needs of archaeologists, and archaeologists were always at the heart of the implementation of the project. As such, ArchAIDE focused on low-tech approaches accessible to the widest range of archaeologists possible, with an emphasis on automating the mundane aspects of existing workflows, while ensuring key points of decision-making within the identification process were left intact. The overarching lesson learned from ArchAIDE was that there is much work to do to create a sufficient amount of well-identified training data to ensure good results from AI-driven workflows in archaeology. ArchAIDE and the Arch-I-Scan Project have shown that AI applications and workflows incorporating automated image recognition using trained neural networks can speed the identification of archaeological pottery. They can better automate the documentation and creation of reliable training data, but to accomplish this, at the kind of scale necessary to make this work more than a proof-of-concept, further work is necessary. The AUTOMated enriched digitisation of Archaeological liThics and cerAmics (AUTOMATA) project is taking many of the lessons learned from ArchAIDE and the Arch-I-Scan Project, while also incorporating robotics to work towards implementing these workflows at scale. This article focuses on the contributions of these types of projects, from a digital stewardship and sustainability perspective, and the important lessons learned regarding how archaeology navigates the use of AI applications in the future.

1. Introduction

The Archaeological Automatic Interpretation and Documentation of cEramics (ArchAIDE) project was an early attempt to automate the process of identifying and documenting archaeological pottery, using automated image recognition, the details of which are also documented in this journal (Anichini *et al*. 2020). The project was meant to address several issues related to these processes: the high volume of pottery found on sites in some parts of the world; the time-consuming and largely analogue workflows associated with making an identification and creating documentation for these types of artefacts; and the general decline in pottery identification as an archaeological specialism (see Cooper and Johnson 2026, this volume, section 5). As the project proposed substantial changes to long-established traditional workflows associated with pottery identification and documentation, after consultation with a range of pottery specialists, an approach was taken that was skeuomorphic, while also carefully considering what aspects of such a workflow could be safely automated, and what aspects required human-in-the-loop intervention. The result of the project was the creation of an app for phones and tablets, and a desktop reference platform using the same underlying dataset. Within the app, which was meant to be used in the field, users would choose one of the three case-study general pottery types (Roman amphorae, Majolica of Montelupo, or *terra sigillata*) and use the built-in camera in the device to photograph a sherd. The user would then deploy the algorithm to identify the specific type in the photograph, returning identifications in order of certainty.

This identification would then be verified by the user, against a digital comparative collection available within the app. Once the user was satisfied with the identification (or entered manually, if no correct match was made), it could be saved into a 'virtual assemblage' along with all the key metadata (both site level and that associated with the pottery type) automatically saved and associated with the image. This also represented significant time saving and ensured the accuracy and consistency of the metadata associated with the artefact. The intention was for this information to be printable for inclusion within a finds bag, and downloadable for analysis and digital preservation. As a proof-of-concept prototype with time-limited funding, the purpose of the ArchAIDE project was not to create a sustainable app or platform, but rather to test how the introduction of digital automation and the incorporation of AI into an analogue workflow might work in practice. The partners worked to address both the technical and human challenges associated with this type of implementation, but the very high number of examples needed to train the network was not sufficiently anticipated. In addition, there was a lack of existing digital comparative data for human-in-the-loop verification of the results returned by the algorithm for Majolica of Montelupo and *terra sigillata*.

To remedy this, time-consuming photography campaigns were undertaken to document thousands of pottery examples in physical collections around Europe to create sufficient training data, and new applications were developed to digitise printed pottery catalogues to allow verification of the pottery type identified by the algorithm within the app. New approaches to creating synthetic training data also had to be developed to boost the accuracy of the algorithm to an acceptable level. All of these issues, and the disproportionate amount of time and effort devoted to their resolution, illustrate the real challenge of applying AI to many tasks in archaeology. The following discussion outlines the contributions working to address this issue made by the ArchAIDE and Arch-I-Scan projects, and the ongoing work of the AUTOMated enriched digitisation of Archaeological liThics and cerAmics (AUTOMATA) project, all of which are currently informing the work of the COST Action *Managing AI in Archaeology* (MAIA).

2. The ArchAIDE Project

The success of the ArchAIDE project relied upon finding sufficient training data to undertake automated image recognition of archaeological pottery, and ensuring that the needs and concerns expressed by pottery specialists were taken into account and built into the

workflow of the ArchAIDE app. The choice of Roman amphorae as one of the case studies for ArchAIDE was based on the fact that there was an existing digital comparative collection that was considered authoritative. This was *Roman Amphorae, a digital resource* (University of Southampton, 2014), a long-established reference collection. It was first published in 2005, with the addition of several new types in 2014. The collection was comprehensive, including hundreds of well-identified amphorae types and sub-types. While not every amphorae type ever documented is represented, the resource stands out due to the depth of information included.

In this digital resource, nearly every type and sub-type includes comparable detailed fields such as Distinctive Features, Date Range, Origin, Distribution, Contents, Comments and how the type correlates with another important comparative resource: the *Centro para el Estudio de la Interdependencia Provincial en la Antigüedad Clásica* (CEIPAC) database. The *Roman Amphorae* archive also includes information on physical characteristics (rim type, shoulder type, height range, etc.), photographs, section drawings using standardised recording and measurement notation, visual and narrative information about petrology, additional specimens, and bibliographic information relating to how the type was defined. The information held within this comparative collection has been made persistently available as a digital archive in perpetuity by the Archaeology Data Service, averaging around 25,000 page views (1,700 unique page views) every month since 2012 (when these statistics were first gathered in their current form).

The success of this single comparative collection, both due to its comprehensiveness and the fact that it has been reliably accessible for over 20 years, made it perfect for use within a project like ArchAIDE, where users would be able to quickly access the information they would need to verify the results of the automated image recognition algorithm within the ArchAIDE app. The resource proved even more useful when the standardised section drawings for each amphora type were turned into 3D models as part of ArchAIDE, and then used to create synthetic training data in the form of 'virtual sherds', crucial to improving the shape-based algorithm (Anichini *et al.* 2021). However, despite the centrality of this dataset, the volume and quality of the synthetic data were not matched by a sufficient number of real images to adequately represent the large number of classes to be identified (more than 200). Unlike *terra sigillata*, one of the other ArchAIDE case studies, where synthetic data were effectively balanced by a higher number of images per class and by a substantially smaller set of archaeological types (i.e. some 50-65 forms), the amphora dataset remained comparatively under-supported. Nonetheless, the creation of this extensive body of synthetic data was far from redundant, as it was then subsequently used to improve the original resource. When the resulting archive was deposited with the Archaeology Data Service by the ArchAIDE project (ArchAIDE consortium 2019), all the 3D models derived from the section drawings were also deposited and then linked with the Amphora archive, meaning the thousands of monthly users of *Roman Amphorae, a digital resource* (University of Southampton 2014) are now able to go directly to the 3D model of the type or sub-type they are researching and interact with it, and the users of the ArchAIDE archive can link through to all the descriptive and diagnostic information about that specific Amphora 3D model when they manipulate or download it. This type of use was never envisioned by the original creators of either archive, but it has become an exemplar of reuse (Wright 2020).

Plans were made to also deposit the thousands of photographs in the ArchAIDE project archive, created during the many campaigns undertaken in physical archives, and used to train the decoration-based automated image recognition algorithm for the Majolica of Montelupo. Unfortunately, this was not possible. While the national heritage authorities in Europe where these campaigns were undertaken allowed the use of the images as training data, the laws regarding the documentation of cultural heritage objects prohibits the open access publication of these images, as they are considered property of the relevant State. This is a significant problem in many European countries, preventing the reuse of data both for the training of algorithms and the creation of digital comparative collections. The

reluctance to allow these types of collections to be created and made open for reuse is understandable. Countries have needed to ensure the safety of their cultural heritage assets for future generations, but projects like ArchAIDE are useful for showing what is possible and why these laws regarding open data may need to be revisited. The challenges and opportunities surrounding this issue are documented in the many papers that make up the *Internet Archaeology: SEADDA Collection* (Wright and Richards 2024). Saving European Archaeology from the Digital Dark Age (SEADDA) was a COST Action that, among many things, used the issues around cultural heritage law and open data learned from the work of ArchAIDE to inform the topics of data stewardship and persistence explored by participants in the COST Action. ArchAIDE was a perfect example of how cultural heritage research generally, and archaeology specifically, was missing out without a more pragmatic approach to open data in many countries. This impact may seem subtle, but without these types of examples, it is very difficult to show the value of open data to those who are responsible for its stewardship.

3. The Arch-I-Scan Project

In contrast, the Arch-I-Scan Project has been able to make a substantial amount of data available for reuse in perpetuity, as it was gathered in the UK, where the creator of an image or scan of an archaeological object holds the copyright, and (in most cases) makes the decision as to how the data should be licensed. The physical archives which provided the artefacts to be documented may have rules and guidance about how their collections should be used (in this case the Arch-I-Scan Project's partner organisations), but there are no overarching laws at the national level stating these cannot be made open access. Starting around the time that ArchAIDE was nearing completion, the Arch-I-Scan Project had similar goals to ArchAIDE regarding using automated image recognition to identify archaeological pottery, but focused much more narrowly on *terra sigillata* imported from Gaul during the Roman period (and some other examples of Roman fineware). *Terra sigillata* was the third pottery type explored by ArchAIDE, but the Arch-I-Scan Project has been able to look much more deeply into how it might be identified.

The photographs created by the Arch-I-Scan Project have now been deposited as an archive with the Archaeology Data Service and it consists of images of pottery from five collections of the project's partner organisations which include the London Museum, Museum of London Archaeology (MOLA), University of Leicester Archaeological Services (ULAS), the Vindolanda Charitable Trust, and the Colchester and Ipswich Museums (CIMS). The dataset from these five collections consists of over 200,000 images (approximately six photographs of each sherd), along with photographs of fabric of the sherds, and more comprehensive photographic documentation of complete vessels (see van Helden and Allison 2026, this volume). The Arch-I-Scan Project has used a similar solution to ArchAIDE by using pottery profiles to create 3D models, which could then be used to create synthetic training data in the form of 'virtual sherds' (see Zheng *et al*. 2026, this volume). The Arch-I-Scan Project now stands as an exemplar for what data management looks like for AI training data.

By making the training data available as well as the results of the research, the basis for how the algorithm was trained is not only more transparent, but the contents of the archive can be reused for training data in the future. The Arch-I-Scan Project's archive also has great potential to provide a foundation for a digital comparative collection for *terra sigillata*, which would certainly be a useful companion to *Roman Amphorae, a digital resource* if someone chose to take the task on. As we make these kinds of data open access and ensure they are available in perpetuity, we understand and can document the impact these resources make. By creating more of them, we can also show the value of this type of research output more generally, which is increasingly seen as useful for academic achievement. The creation of the Arch-I-Scan Project's archive was a major undertaking by the project team, and it is important to acknowledge why this work is important, and how best to demonstrate it.

4. The AUTOMATA Project

The primary follow-on project from ArchAIDE and the Arch-I-Scan Project is AUTOMated enriched digitisation of Archaeological liThics and cerAmics (AUTOMATA). AUTOMATA is taking what was learned from ArchAIDE and the Arch-I-Scan Project, in terms of the application of automated image recognition to archaeological pottery, and expanding it to include lithics, automated analyses and robotics. AUTOMATA is one of the three initial data-providing projects for the European Cloud for Heritage OpEn Science (ECHOES) as part of the major European infrastructure investment, the European Collaborative Cloud for Cultural Heritage (ECCCH). The goal of AUTOMATA remains much the same as the previous projects: to find ways to speed up very time-consuming tasks in archaeology while determining where and how human-in-the-loop interventions need to be implemented in documentation workflows. In this instance, the documentation created by the project will not only be archived with the Archaeology Data Service, but also made cross-searchable from the ECHOES infrastructure (and the ARIADNE Portal for archaeological data).

AUTOMATA is ambitious, and is working to create a system that can use newly developed 'soft touch' robotics in the form of a robotic arm that will be able to pick up sherds of pottery and lithic tools without damaging them. It will hold the artefact, allowing it to be documented using photogrammetry and laser scanning, and subjected to a range of analyses such as spectroscopy and X-ray fluorescence. The tools used for documentation and analyses will include existing AI methods to improve the digitisation process and the quality of the data being captured, but it will also create training data under robust and controlled conditions for use by AI algorithms. AUTOMATA has the potential to build a workflow that creates larger datasets that are necessary for creating training data for AI applications, and to create the data in such a way that they are automatically optimised for that purpose. Many of the physical practicalities of digitising pottery for use by AI (such as, at what angle must the sherd be photographed or scanned, or how many photographs must be taken?) were understood through the work of ArchAIDE and the Arch-I-Scan Project. These practical insights are now being used to inform the automated processes for the robotic arm in AUTOMATA.

Another key distinction between ArchAIDE and AUTOMATA is the requirement to make the data generated by the project open access and persistently available for reuse. Under the funding requirements of ECCCH, the purpose of the project is to create data that can be accessed via the ECHOES infrastructure. The partners providing access to the artefacts to be used by AUTOMATA had to agree that the data could be made open, and that the Archaeology Data Service would be responsible for providing both the long-term preservation of the data and supplying the metadata from the archive to ECHOES. Even with these assurances in place, the project may still encounter obstacles regarding what can be made open access under the cultural heritage laws of the countries holding the artefacts. Experience has shown that interpretation of these laws, rules and norms can shift when open-data implementation becomes imminent and final decisions must be made. Having positive exemplars to point to, such as ArchAIDE and the Arch-I-Scan Project, is going to be very helpful for the work of AUTOMATA.

5. COST Action MAIA

Another important outcome from the lessons learned from ArchAIDE and the Arch-I-Scan Project, and a direct contributor to the approach of AUTOMATA, is the COST Action Managing AI in Archaeology (MAIA). COST Actions are networking projects that span four years, and members can join at any time. The interest in MAIA has been extremely high relative to other COST Actions, with over 350 people from 34 different countries choosing to join in the first year alone. This can only be a testament to the level of engagement archaeologists wish to have with AI, and of AI specialists who wish to engage with archaeological data. Despite high interest, there is no comprehensive understanding of the current state-of-the-art in AI and archaeology; documenting how AI is presently applied to

archaeological data is therefore a primary focus of MAIA. This may prove difficult, however, as AI adoption is accelerating so rapidly that trying to compile information on its use and application even within a single discipline has become daunting.

The most direct connection between the needs identified by ArchAIDE and the Arch-I-Scan Project is MAIA's focus on determining how to achieve reliable results when applying AI to archaeological data, and how to validate those data through human-in-the-loop interventions. This entails exploring how to generate more training data, how to ensure the data is open access so that others can train on them, and how to expand digital comparative collections (Anichini and Gattiglia 2022). In other words, how we can create more open-access, persistently available data that meet the needs of AI applied to archaeological data. Much of this work is focused on raising awareness about these needs (Gattiglia 2025), while also trying to identify the resources that could (or should) be made available for these purposes.

Finally, MAIA members are working to establish how AI *should* be applied to archaeological data. This includes exploring the potential of AI applications to answer archaeological research questions, how such workflows might be implemented, and what ethical implications they carry. Although open-access archaeological data have undoubtedly been absorbed into commercial platforms regardless of stated terms of use and access, archaeologists are not powerless to shape the future of AI in their field. By developing fluency in both the challenges and opportunities offered by AI, archaeologists can actively lead the conversation - defining how our data should be used while maximising the benefits AI has to offer.

6. Conclusion

The interconnected projects and ongoing work described here are meant to show that the use of AI in archaeology depends as much, if not more, on the creation of open-access, persistently available data as on the technical development of algorithms. As AI becomes an ubiquitous part of our research (and general) lives, it is important that we also put some of our time and effort into ensuring that the data we need to ensure algorithms give good quality results are available not only to us, but to others. Only by working together can humanities-based disciplines like archaeology build the training datasets large enough to meet our needs. Simultaneously, we must invest in the creation of digital comparative collections to ensure we can leverage AI-driven automation while still being able to verify the results. The landscape of AI applied to archaeological data is changing rapidly, but these are fundamental issues that need to be addressed to ensure the results we produce are of high quality, and can be trusted as part of our research.

Bibliography

Allison, P. and van Helden, D. 2026 *The Arch-I-Scan Project. 2019-2023* [dataset], York: Archaeology Data Service [distributor]. https://doi.org/10.5284/1138130

Anichini, F., Banterle, F., Buxeda i Garrigós, J., Callieri, M., Dershowitz, N., Dubbini, N., Diaz, D.L., Evans, T., Gattiglia, G., Green, K. and Gualandi, M.L. 2020 'Developing the ArchAIDE application: a digital workflow for identifying, organising and sharing archaeological pottery using automated image recognition, *Internet Archaeology* **52**. https://doi.org/10.11141/ia.52.7

Anichini, F., Dershowitz, N., Dubbini, N., Gattiglia, G., Itkin, B. and Wolf, L. 2021 'The automatic recognition of ceramics from only one photo: The ArchAIDE app', *Journal of Archaeological Science: Reports* **36**, 102788. https://doi.org/10.1016/j.jasrep.2020.102788

Anichini, F. and Gattiglia, G. 2022 'Reflecting on artificial intelligence and archaeology: the ArchAIDE perspective', *European Journal of Post-Classical Archaeologies* **12**, 69-86. https://www.postclassical.it/PCA_Vol.12.html

ArchAIDE consortium 2019 *ARCHAIDE Portal for Publications and Outputs* [data-set], York: Archaeology Data Service [distributor]. https://doi.org/10.5284/1050896

Cooper, N.J. and Johnson, E. 2026 'It Cuts Both Ways! Reflections on the Contribution of the Pottery Specialist to the Arch-I-Scan Project and the Benefits of Being Involved', *Internet Archaeology* **74**. https://doi.org/10.11141/ia.74.5

Gattiglia, G. 2025 'Managing artificial intelligence in archeology. an overview', *Journal of Cultural Heritage* **71**, 225-233. https://doi.org/10.1016/j.culher.2024.11.020

University of Southampton 2014 *Roman Amphorae: a digital resource* [data-set], York: Archaeology Data Service [distributor]. https://doi.org/10.5284/1028192

van Helden, D.P. and Allison, P.M. 2026 'The Arch-I-Scan Project's recording and image preparation processes: development and refinements', *Internet Archaeology* **74**. https://doi.org/10.11141/ia.74.3

Wright, H. 2020 'Digital Archives, Open Data and Unexpected Legacies', *The Archaeologist*, **110**, 16–18.

Wright, H. and Richards, J. D. (eds) 2024 'Internet Archaeology: SEADDA Collection', *Internet Archaeology* [Online]. https://doi.org/10.11141/ia.seadda.1 [Last accessed: 3 Sept 2026]

Zheng, Z., van Helden, D.P., Allison, P.M., Tyukin, I.Y. and Tyukina, T.A. 2026 'Developing Usable Datasets for AI training: Real and simulated sherds', *Internet Archaeology* **74**. https://doi.org/10.11141/ia.74.7