GA – PROJECT NUMBER:	**101158046**
PROJECT ACRONYM:	AUTOMATA
PROJECT TITLE:	AUTOmated enriched digitisation of Archaeological liThics and cerAmics
CALL/TOPIC:	HORIZON-CL2-2023-HERITAGE-ECCCH-01-02
TYPE OF ACTION	HORIZON RIA
PRINCIPAL INVESTIGATOR	Prof Gabriele Gattiglia, UNIPI
TEL:	+39 050 2215228
E-MAIL:	gabriele.gattiglia@unipi.it

This project has received funding from the European Union's HORIZON RIA research and innovation programme under grant agreement N. 101158046

D 10.1 Data Management Plan

Version: 1.0

Revision: first release

Work Package: **10 - Data Management and Curation**
Lead Author (Org): **Holly Wright (UoY)**
Contributing Author(s) (Org): **Gabriele, Gattiglia (UNIPI)**
Due Date: **M6**
Date: **28/02/2025**

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	**X**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Description
.5	11/2/25	Holly Wright	Initial draft completed
.8	24/2/25	Gabriele Gattiglia	Text on data specifics added
1.0	27/2/25	Holly Wright	Deliverable finalised

Disclaimer

Abbreviations

WP: Work package

M: Month

UNIPI: Universita di Pisa

UBM: Universite Bordeaux Montaigne

UoY: University of York

INRAP Institut National de Recherches Archeologiques

AMZ: Arheoloski Muzej u Zagrebu

QB: QBrobotics Srl

HUJ: The Hebrew University of Jerusalem

MIN: Miningful srls

KCL: King's College London

IIT: Fondazione Istituto Italiano di Tecnologia

UB: Universitat de Barcelona

CL: Culture Lab

Overview

Data management presents a significant challenge, particularly due to the data-intensive nature of 3D models enriched with archaeometric information. While chemical data is relatively lightweight, handling 3D data requires an eco-responsible approach. This means digitising what is necessary to address the specific scientific question rather than uniformly capturing an entire object at the same resolution and precision. Instead, the 3D digitisation process should be adapted to focus on different parts of the object based on the relevance of the information or analyses those areas can provide.

This also entails exploring data compression techniques for 3D models to manage storage without sacrificing accuracy, supporting easy retrieval and processing. Moreover, the system should include a digital ID system to ensure each artefact can be easily found post-digitisation, facilitating further analysis or re-evaluation.

Data will be derived from the following sources:

- **Archaeological excavations**: Ceramic and lithic artefacts are recovered during archaeological excavations. The quantity and quality of the materials collected depend on excavation strategies and the characteristics of the deposit. The materials are labelled with context numbers and stored in transparent bags.
- **3D digitisation**: 3D models are created using various data acquisition technologies, such as photogrammetry, laser scanning, and structured light. These technologies capture 3D data in different ways but ultimately all produce a point cloud. 3D models are used for visualisation, morphometric, and spatial analysis.
- **Archaeometric analyses**: The data derive from physical, chemical, and mineralogical archaeometric analyses conducted on ceramic and lithic artefacts. These analyses provide detailed information on the composition, provenance, and technological processes of the artefacts. Analyses can be destructive or non-destructive, invasive or non-invasive. Techniques such as Raman spectroscopy and portable X-ray fluorescence (p-XRF) are used for non-invasive analysis.
- **Hyperspectral imaging (HSI)**: Hyperspectral cameras capture the reflectance spectrum at each pixel of an image, revealing the wavelengths of visible and near-infrared light reflected by the object. HSI data can be used to identify specific materials and map decorated surfaces.
- **Contextual data and metadata**: The data also include contextual information on the artefacts, such as their find location, associations with other artefacts, and stratigraphic context. Metadata provide a structured way to document and interpret archaeological finds, facilitating their integration into broader analytical frameworks.
- **3D Referenced Information System (RIS3D)**: Integrates 3D models with detailed analytical data, allowing spatial exploration and querying of the data within a 3D interface. The data are stored in a PostgreSQL relational database and in JSON format.

We have specified the formats of raw data obtained from the various sensors in D2.1 in Table 2, Section 2.2.3 (Sensors specifications and requirements), as well as in Section 2.3.1 (Automating the Referenced Information System in 3D (RIS3D) process and software requirements), where we discuss the data that are input into RIS3D. These include:

- **iPhone LiDAR and photogrammetric scanning**: generates point cloud data exportable in standard 3D formats such as .obj, .usd, or .ply.
- **Polymetric PT-M4 3D Scanner**: produces visual 3D models and point cloud coordinate data. Exports geometry as .wrl/.vrml, .stl, .obj, and .ply files.
- **IQ Hyperspectral Camera**: produces reflectance spectra that can be processed into conventional images (RGB or False-Colour Infrared) or more advanced statistical treatments. Spectral data are in .csv, .tiff, and .hdr formats.
- **Vanta p-XRF (Olympus)**: common formats include .csv, .xls, and .txt, suitable for analysis in various software applications. Produces two spectra (one per beam).
- **Bravo Raman (Bruker)**: generates Raman spectra in text file format (.dpt), .dat, or opus (.0).
- **i-Raman Plus 785H (Metrohm - BWTek)**: spectral data in text file format (e.g., .txt).

RIS3D (Referenced Information System in 3D):

- Data are stored in a PostgreSQL relational database.
- Data are stored in the database in JSON format.
- JSON supports a wide range of data types, including text annotations, measurements, external links, images, files, dates, and boolean values.

As set out in Section 2 of D2.1: it is important to highlight a distinction between the traditional users who typically conduct artefact analysis and digitisation and the broader group of stakeholders identified by the AUTOMATA project. Traditionally, these tasks are carried out by specialised professionals who invest many hours in the manual or semi-manual creation of 3D models and archaeometric analysis of objects. In contrast, the AUTOMATA project aims to automate parts of this process, thus expanding the system's potential user base beyond these specialists.

By introducing automation, AUTOMATA broadens access to high-quality digitisation and analysis, allowing a more diverse range of stakeholders to engage with the system. Key stakeholders for AUTOMATA include academics, museum professionals, heritage supervisory bodies, students, and professional archaeologists. Each group will interact with the project's outputs in ways tailored to their needs. Academics will use the data for research, education, and outreach, while the system also holds significant potential for teaching purposes, particularly in demonstrating how movable instruments can generate valuable insights into cultural heritage. Ministries and heritage bodies will utilise the data for heritage protection, preservation initiatives, and public communication, supporting responsible cultural heritage stewardship. In museums, the digitised data and instruments will enable curators, archaeologists and researchers to improve artefact processing, inventory management, analysis, publication, and educational outreach.

AUTOMATA Project partners will be responsible for the safe handling, including regular backups, of their project data while it is under development. The University of Turin has provided partners with a dedicated institutional Google Drive for the project. Regular backups will be performed using physical hard disks.

Project data in its final form will be archived with the Archaeology Data Service (ADS) where ownership of the data can be attributed to the project consortium.

The ADS accepts all data types and formats that will be produced by the AUTOMATA Project, including databases, spreadsheets, text files, XML datasets, images, 3D models, audio and video. While it is difficult to estimate the amount of data that will be produced by the project that will be suitable for deposit, the ADS has access to a petabyte of storage, so whatever amount of data is produced can be accommodated.

The ADS is an advocate for the FAIR principles for data stewardship. The ADS actively work to assess how the datasets it curates can be more fully compliant with the FAIR principles, but the following sections describe the current FAIRness of all data archived by the ADS, and therefore the level of FAIRness that will be implemented for the AUTOMATA archive.

1 FAIR Data

Project data in its final form will be archived with the Archaeology Data Service (ADS) where ownership of the data can be attributed to the project consortium. The ADS is always working to improve the FAIRness of the data it holds. Currently, all data archived with the ADS complies with the FAIR Principles in the following specific ways, but additional improvements may be implemented by the time the data generated by AUTOMATA will be ready for deposit.

1.1 Making data findable

F1. (Meta)data are assigned a globally unique and persistent identifier.

- The ADS uses Digital Object Identifier (DOIs) persistent identifiers for all collections.
- The ADS supports the use of ORCID IDs.
- The ADS supports the use of WikiData Q Codes

For a fuller discussion of the Archaeology Data Service (ADS) metadata and the use of persistent identifiers see the ADS Metadata policy and procedures pages.

F2. Data are described with rich metadata (defined by R1 below).

- All ADS resources are documented using the Dublin Core Metadata Element Set (DCMES) plus Dublin Core Metadata Initiative (DCMI) recommended qualifiers.
- The ADS also provides rich qualitative and technical metadata for all digital objects. These are repository specific metadata requirements, derived from domain-specific community standards (i.e. Guides to Good Practice, see also R1.3 below).
- All metadata is displayed alongside data, with technical metadata downloadable in open formats.

F3. Metadata clearly and explicitly include the identifier of the data they describe.

- All persistent identifiers for ADS collections are clearly displayed, alongside data, within each archive interface.
- The ADS supports the use of additional or supplemental identifiers relating to the dataset that link to external repositories, agencies or resources. This includes identifiers for physical, as well as digital, collections.

F4. (Meta)data are registered or indexed in a searchable resource.

ADS datasets are findable through the repositories own indexes and catalogues.

- ArchSearch
- Archives
- ADS Library

ADS collections are also available through external catalogues and resources, including:

1. Heritage Gateway
2. DataCite
3. the Keepers Registry
4. Natural Environment Research Council (NERC) data discovery portal
5. ARIADNEPlus Portal

6. Marine Environmental Data and Information Network (MEDIN) data portal
7. Europeana

ADS catalogues and indexes are searchable and harvestable through a series of OAI-PMH targets, and as linked open data using a SPARQL query web interface.

1.2 Making data accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.

- All ADS datasets utilise the HTTPS protocol to ensure free and open access to resources and to facilitate data retrieval.
- In rare instances, where discrete data objects are too large to support easy exchange using HTTPS, the ADS makes data available 'on request' using free and open exchange services (e.g. University of York DropOff Service).

A1.1 The protocol is open, free, and universally implementable.

- The ADS uses the HTTPS protocol for the sharing of resources and transfer of datasets. This is widely supported, open, and freely available.
- The repository utilises open and free file-sharing services where files or datasets are too large for easy exchange using HTTPS. Typically the ADS utilises the open and free University of York DropOff Service to share data when this is necessary.

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.

- The use of HTTPS provides authentication of the ADS website, and ensures the protection of the privacy and integrity of disseminated data. The repository ensures that all server-side digital certificates are current and up to date.
-

A2. Metadata are accessible, even when the data are no longer available.

- As an accredited digital repository the ADS supports long-term preservation and access of its holdings, consequently all datasets and metadata are maintained in perpetuity.
- The ADS maintains a clear Appraisal and Deaccession Policy which outlines current practice for datasets removed from the archives holdings. In such instances the ADS is committed to supporting identifiers (i.e. DOIs), maintaining resource discovery metadata, and updating current information on resources.

1.3 Making data interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- All resource discovery metadata is made available using a qualified Dublin Core in RDF/XML through the ADS Linked Data repository
- External services also consume and disseminate metadata.

I2. (Meta)data use vocabularies that follow FAIR principles.

The ADS uses a variety of sustainable, open vocabularies to qualitatively classify and identify resources and datasets, including:

- Heritage Data vocabularies, including those provided by the Forum on Information Standards in Heritage (FISH), Historic England (HE), Historic Environment Scotland (HES), and the Royal Commission on Ancient & Historical Monuments of Wales (RCAHMW)
- Library of Congress Subject Headings (LCSH)
- Marine Environmental Data and Information Network (MEDIN)
- Getty Thesaurus of Geographic Names (TGN)

The ADS also utilises recognised technical vocabularies to denote and categorise preservation activities.

- PREservation Metadata: Implementation Strategies (PREMIS)
- Getty metadata types (Baca 2016)

I3. (Meta)data include qualified references to other (meta)data.

- The ADS supports the qualified referencing with and between publications, datasets and resources. Where available the repository uses sustainable referencing, e.g. DOIs.

For a wider discussion on the vocabularies used in ADS metadata see the ADS Strategy and Standards page.

1.4 Making data reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

- All ADS resources have clearly defined terms of access and reuse within each collection interface, and within metadata records distributed by the ADS or externally. Typically, data is disseminated under the terms of Attribution 4.0 International (CC BY 4.0), but data may also be disseminated under other forms of Creative Commons (see also the ADS Terms of Use and Access to Data).

R1.2. (Meta)data are associated with detailed provenance.

- The ADS provides detailed provenance metadata for all data. At a collection level this is clearly expressed in the archive interface and discovery metadata, but also at a file level within the technical metadata disseminated alongside the data.

R1.3. (Meta)data meet domain-relevant community standards.

- The ADS utilises a qualified Dublin Core metadata standard for all collection level metadata (noted above). The repository also uses standardised templates to ensure metadata consistency. All data must be accompanied by appropriate, file specific 'technical' metadata, this is derived from recognised community standards (Guides to Good Practice) to ensure consistency. All (meta)data is accepted, preserved and disseminated in sustainable, open formats. These are expressed in the ADS Instructions for Depositors and the ADS Policy and Procedures. The repository employs appropriate vocabularies to qualitatively describe datasets (noted above) and document preservation actions.

2 Other Research Outputs

Any research outputs not published elsewhere can be included as part of the AUTOMATA project archive, and where published or made available elsewhere will follow the previously mentioned FAIR sub-Principle:

I3. (Meta)data include qualified references to other (meta)data.

- The ADS supports the qualified referencing with and between publications, datasets and resources. Where available the repository uses sustainable referencing, e.g. DOIs.

3 Allocation of Resources

Project data in its final form will be archived with the Archaeology Data Service (ADS) where ownership of the data can be attributed to the project consortium. The cost of archiving the data with the ADS has been allocated within the AUTOMATA budget.

4 Data Security

The ADS is the leading accredited digital repository for archaeology and heritage data. Founded in 1996, the core activity of the ADS is long-term digital preservation. The ADS follows a policy of active data management and curation to ensure the integrity, reliability and accessibility in perpetuity of the data they hold. All resources archived with the ADS are Open Access and delivered through a website to facilitate re-use by the heritage sector and wider community. The ADS is a world leader in promoting good practice in the use of digital data in archaeology, providing technical advice to the research community and leading a wide range of research projects.

To ensure the highest level of compliance and service provision, and build trust with its designated community, the ADS has submitted itself to formal review undertaken by external service providers and agencies. Consequently, the ADS has been awarded the Data Seal of Approval and its replacement, the CoreTrustSeal, is a regular member of the World Data System (WDS), holding the WDS Certification of Trustworthy Digital Repository. Further details related to data security can be found in the Archaeology Data Service Preservation Policy.

As a founding partner of the ARIADNE RI, the ADS now aggregates the resource discovery metadata for all archives within the ARIADNE Portal, making them cross-searchable alongside over four million other archaeological resources. This means the AUTOMATA data deposited for archiving with the ADS will also be discoverable via the ARIADNE Portal.

5 Ethics

As stated in the AUTOMATA application:

Non-EU countries: non-EU countries will not have access to any kind of personal data or sensitive material during the project but will offer advice on certain technical steps in setting up the system. A part of data

acquisition will be carried out in non-EU countries, but historical materials will not be transported out of the country and the activities will be carried out according to the current laws in force in the state of Israel. The Computational Archaeology Laboratory at The Hebrew University of Jerusalem (HUJI) will actively participate in WP2 and 9 and in Task 5.2 and will lead Task 8.4.

Concerning AI: Aware of the possible risks associated with the use of AI systems, AI technical robustness is inherently built into our research programme. To deal with any failures, inaccuracies and errors, we have committed to a design with guaranteed continual learning capabilities which builds on our unique expertise and track record in that area. We also will capitalise on the state-of-the-art explainable AI tools to highlight and identify potential risks of misclassification and erroneous decisions. To ensure that all results are technically trustworthy, we have employed careful calibration tasks, enabling the assignment and reporting of a confidence interval for decisions produced by the system. AI systems will be involved in WP9, tasks 4.2 and 5.4.

The data collection that will be carried out with the Israeli partner will comply with the applicable local legislation. We also declare that the objects will not deteriorate or be affected in any way as this is a non-invasive analysis. This treatment of historical materials is fully in accordance with the applicable European cultural heritage laws. Data collection at HUJI will be carried out in WP8.

Concerning AI: The Project is based on the Ethics Guidelines for Trustworthy AI by an Independent High-Level Expert Group on Artificial Intelligence set up by EU Commission (HLEG, 2019). Therefore, it promotes the development and use of a lawful, ethical and robust AI system. A Trustworthy AI assessment list is adopted when developing, deploying, or using AI systems and adapted to the specific use case in which the system is being applied. In particular, since this is a digitisation system, special attention is paid to points 3) Privacy and data governance (respect for privacy and data protection, quality and integrity of data, access to data), and 4) Transparency (traceability, explainability, communication).