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 5.2 3D Database

Version: 1.0

Revision: first release

Work Package: 5 - Technologies for enriched digitisation
Lead Author (Org): Arthur Leck (UBM)
Contributing Author(s) (Org): Rémy Chapoulie (UBM), Martina Naso (UNIPI), Romain Pacanowski (INRIA), Clément Joubert (INRIA), Nevio Dubbini (MIN), Daniel Van Helden (KCL), Sarah Tournon (UBM), Mikaël Rouca (UBM), Pascal Mora (UBM), Heeli Chaya Schechter (HUJ), Gabriele Gattiglia (UNIPI)
Due Date: M16
Date: 31/12/2025

Dissemination Level		
P	**Public**	**X**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Description
0.1	26/09/2025	Arthur Leck (UBM), Martina Naso (UNIPI)	Structure and draft content
0.2	12/11/2025	Martina Naso (UNIPI)	Draft content added
0.3	09/12/2025	Clément Joubert (INRIA)	Draft content for 2.1
0.4	10/12/2025	Romain Pacanowski (INRIA)	Write content for 2.1
0.5	11/12/2025	Martina Naso (UNIPI)	Draft content added
0.6	15/12/2025	Nevio Dubbini (MIN)	Draft content added
0.7	16/12/2025	Daniel Van Helden (KCL)	Draft content added
0.8	17/12/2025	Arthur Leck (UBM), Sarah Tournon (UBM), Mikaël Rouca (UBM), Pascal Mora (UBM)	Draft content added
0.9	18/12/2025	Rémy Chapoulie (UBM)	Draft content suggestions
0.10	22/12/2025	Arthur Leck (UBM)	Draft content added
0.11	24/12/2025	Heeli Chaya Schechter (HUJ)	Draft content suggestions
0.12	29/12/2025	Arthur Leck (UBM), Martina Naso (UNIPI), Nevio Dubbini (MIN)	Revision of the document
0.13	05/01/2026	Gabriele Gattiglia (UNIPI)	Draft content and suggestions added
0.14	08/01/2026	Arthur Leck (UBM)	Draft content added
0.15	13/01/2026	Arthur Leck (UBM), Gabriele Gattiglia (UNIPI)	Final revision of the document

Disclaimer

Abbreviations **4**
Executive summary **5**
1 Introduction **6**
2 Test materials **8**
3 Photogrammetry Workflow **10**
3.1 Description of the acquisition process 10
3.1.1 Acquisition process at INRIA 10
3.1.2 Acquisition process at Archeovision (UBM) 12
3.1.2.1 Close-range photogrammetry 12
3.1.2.2 Resolution tests using focus stacking 14
3.2 Role of AI in photogrammetry 16
4 Archaeometric Data Workflow **18**
4.1 Hyperspectral Imaging 18
4.2 Portable XRF 22
4.3 Raman Spectrometry 24
4.4 Role of AI in the archaeometric workflow 25
5 Data Management and Database Infrastructure **27**
5.1 Storage 32
6 Conclusions and future development **33**
References **34**

Abbreviations

WP: Work Package

M: Month

UNIPI: Università di Pisa

UBM: Université Bordeaux Montaigne

UoY: University of York

INRAP: Institut National de Recherches Archéologiques

INRIA: Institut national de recherche en sciences et technologies du numérique

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

Executive summary

Deliverable 5.2 presents an initial implementation of the AUTOMATA 3D database, developed to store, organise, and manage 3D appearance and archaeometric data acquired throughout the project. Here, “3D database” is used in an operational sense to denote a hybrid data backbone: a structured repository on the ArcheoGRID sandbox for storing and organising raw and processed digital assets, together with a metadata layer implemented in PostgreSQL (relational core) with JSON/JSONB structures to represent technique-specific parameters. This implementation provides persistent identifiers and traceability across acquisition and processing steps, and is designed to support ingestion into the RIS3D integration (D5.3), where spatially anchored object-level data will be managed in an integrated environment and queried. The core output of this deliverable is the database itself, hosted on the ArcheoGRID platform (https://www-dev.archeogrid.fr/project/12527), and the metadata model (https://www-dev.archeogrid.fr/viewer/12697_137584?format=hhdd). The document provides a concise account of the acquisition workflows, data organisation, and methodological choices that guided its construction. The work builds upon the standards and guidelines defined in Deliverables 5.1 (Ontology and Metadata Scheme for Enriched Digitisation) and 10.1 (Data Management Plan). Section 2 introduces the test artefacts selected for this phase, including ceramic and lithic artefacts chosen to validate both appearance and archaeometric acquisition workflows. The section details the provenance, typology, and selection criteria of the samples, establishing the reference corpus used to populate the database's initial version. Sections 3 and 4 focus on the acquisition workflows for appearance data and archaeometric data, respectively. Section 3 documents the photogrammetric workflows used to generate 3D models, describing acquisition setups, experimental protocols, and resolution tests carried out at INRIA and Archeovision (UBM). It also discusses the challenges posed by thin, reflective, or small artefacts and outlines the supporting role of AI in photogrammetry, including image segmentation, data validation, and model simplification. Section 4 details the archaeometric workflows, covering hyperspectral imaging, portable XRF, and Raman spectrometry. These sections explain the structure of the raw and processed outputs as well as the use of a specific software for data normalisation and exploratory analysis. Section 5 describes the data management strategy and database infrastructure. It presents the organisation of raw and processed data within the ArcheoGRID sandbox, the metadata architecture combining relational and JSON-based components, and the mechanisms ensuring traceability and interoperability between 3D models and analytical datasets. This section also outlines storage considerations and strategies for managing large data volumes while preserving reproducibility and long-term usability. Finally, Section 6 summarises the current state of the database and outlines future developments. It positions Deliverable 5.2 as a foundational step toward integrating enriched 3D data within the Referenced Information System in 3D (RIS3D), to be delivered in Task 5.3, and toward the progressive automation of acquisition, metadata generation, and enriched 3D model production in subsequent tasks and work packages.

1 Introduction

This deliverable concerns the data backbone supporting the initial UBM acquisitions of 3D appearance and archaeometric data. This document is a supplement explaining the various stages that led to the construction of this infrastructure. It documents a hybrid infrastructure rather than a single monolithic relational schema: (i) a structured file repository hosted on the ArcheoGRID sandbox, used to store the digital assets produced by photogrammetry and archaeometric pipelines (raw and processed) according to a consistent directory logic; and (ii) an associated metadata layer implemented in PostgreSQL, combining a relational core with JSON/JSONB structures to represent heterogeneous, technique-specific parameters. Together, these components ensure persistent identifiers, provenance and traceability, and provide the controlled input required for the next-stage RIS3D implementation (D5.3), where the integration of analytical measurements with the 3D model (including spatial anchoring) will be operationalised.

Deliverable 5.2 marks a critical step in developing a robust data management backbone that will serve as the foundation for the systematic management of all collected data during the AUTOMATA project. The main objective of this deliverable is to establish a well-structured data management infrastructure that accommodates different levels of data processing, including both raw and processed versions, and supports both immediate research needs and long-term accessibility. The main data types and file formats generated by the pipelines are described in Section 4 and are referenced in Section 5 in terms of storage, registration, and linkage to metadata records.

A key aspect of this phase is the careful acquisition of appearance and archaeometric data, combined with a clear and detailed description of the methodologies used. In addition, the variety of objects and techniques tested at this stage has been used to derive practical requirements for the data backbone, informing how assets, parameters, and provenance information are structured and linked in the database infrastructure. The focus is to create datasets that serve multiple purposes: validating the data acquisition process, providing a benchmark for quality control, and laying the groundwork for subsequent model enrichment. By collecting archaeometric information independently at this stage, the project ensures that all relevant material characteristics are captured systematically, even before they are incorporated into the final 3D models.

Normally, acquisition protocols are documented through metadata and integrated directly into the database. This approach works well for protocols that are relatively stable, allowing metadata to be generated automatically. In this project, however, the protocols and tests varied considerably, and the metadata could not capture all the experimental conditions. For this reason, the different acquisition protocols and tests are described explicitly in this deliverable. These descriptions represent work in progress and are intended to provide a clear record of the procedures used, rather than to be archived in rigid metadata formats. Nevertheless, the metadata file has been developed and is already available in the database (see below), ready for use in the implementation of RIS3D.

This work builds directly on the framework established in Deliverables 10.1 (Data Management Plan) and 5.1 (Ontology and Metadata Scheme for Enriched Digitisation), which define the initial standards for data collection, processing, and organisation. In line with these guidelines,

Deliverable 5.2 consolidates raw, processed and compressed 3D and archaeometric data in a reference dataset fully compatible with forthcoming tasks. In particular, it has to align with the Referenced Information System in the 3D (RIS3D) platform, which will be available as part of Deliverable 5.3 (Reference Enriched 3D Data). This platform will enable the automatic linking of an artefact's chemical and physical characteristics to its 3D model, recording both the spatial coordinates of the measurements and their respective values, and allowing archaeometric data to be precisely located on each vertex of the 3D mesh.

2 Test materials

For Tasks 5.2 (*Creation of 3D models for pottery and lithics*) and 5.3 (*Creation of the enriched 3D model*), a selection of ceramic and lithic fragments was made to evaluate the appearance and archaeometric acquisition and processing workflows.

A total of 300 ceramic fragments were provided by INRAP, organised into three distinct groups:

- Lot 1, i.e. 114 fragments from a Roman-period deposit in Rennes (France), combining waste from a potter's workshop and domestic refuse, including locally produced wares, imitations, and Roman terra sigillata.
- Lot 2, i.e. 147 fragments of Roman terra sigillata originating from different production workshops but recovered from a single site.
- Lot 3, i.e. 39 Bronze Age sherds, comprising domestic pottery and vessels associated with salt production.

For the purpose of this deliverable, a subset of 20 sigillata fragments has been selected from Lots 1 and 2 to initiate the population of the AUTOMATA database. The selection was carried out using randomised sampling, ensuring variability in morphology and size and allowing the workflows for appearance and archaeometric data to be tested across different conditions. The resulting sample includes 3 fragments from Lot 1, all assigned to the sigillata Lezoux production group (typological form Drag. 37), and 17 fragments from Lot 2, representing a broader range of forms and production centres typical of sigillata (fig. 1). No samples were selected from Lot 3, as this deliverable prioritised artefacts' morphological complexity relevant to 3D modelling and focused, at this initial stage, on a single ceramic category (sigillata), not present in the Bronze Age Lot 3. This subset provides a heterogeneous yet coherent starting point for validating the acquisition, processing and integration procedures required for the AUTOMATA database.

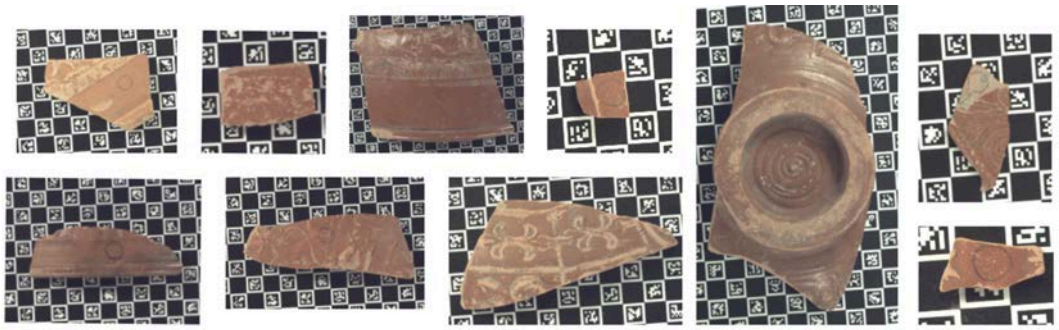

Fig. 1 - Example of 10 sigillata fragments from Lots 1 and 2 included in the selected subset.

In parallel with the ceramic samples, a set of 22 lithic artefacts was selected to represent a wide range of sizes, shapes, and raw materials, enabling testing of a variety of acquisition parameters.
Part of the samples was selected from those available at the Archéosciences Bordeaux laboratory. They come from experimental archaeology studies or from archaeological objects collected during surface surveys carried out in various archaeological operations. These samples have a unique laboratory identification number (ex. BDX27801).

A second part of the samples was provided by INRAP and corresponds to the results of experimental archaeology. They consist of an almost complete refitting of a Levallois reduction sequence. The samples are named from R37_1 to R37_10.

Tab. 1 - List of the selected lithic fragments with their unique laboratory identifiers, materials, acquisition contexts (experimental or archaeological), provenance (site), and a brief description. It also indicates which analyses have been performed yet for each sample, including 3D modelling, HSI, pXRF, and Raman spectrometry.

SampleName	Material	Type	Site	Description	3D	HSI	pXRF	Raman
BDX27801	Ignimbrite	experimental	Stagnu plateau	BlocStagnu ExpeJV Armature	yes	yes	yes	yes
BDX27802	Ignimbrite	experimental	Stagnu plateau	BlocStagnu ExpeJV GrdEclat	yes	yes	yes	yes
BDX27803	Obsidian	archeological	Monti Barbatu, Olmetu	MB17 HS Lamelle Obsi1	no	yes	yes	yes
BDX27804	Obsidian	archeological	Monti Barbatu, Olmetu	MB17 HS Lamelle Obsi2	yes	yes	yes	yes
BDX27805	Rhyolite	experimental	Fangu valley	RhyoVerte ExpeJV PercuInd L1	no	yes	yes	yes
BDX27806	Rhyolite	experimental	Fangu valley	RhyoVerte ExpeJV PercuInd L2	no	yes	yes	yes
BDX27807	Rhyolite	experimental	Fangu valley	RhyoVerte ExpeJV PercuInd L3	no	yes	yes	yes
BDX27808	Rhyolite	experimental	Fangu valley	RhyoVerte ExpeJV PercuInd nucleus	yes	yes	yes	yes
BDX27809	Chert	archeological	around Bergerac	Silex prospe Berg GrosEclat	yes	yes	yes	yes
BDX27810	Chert	archeological	I Casteddi, Tavera	TAV18 SILEX HS EP. S.E	yes	yes	yes	yes
BDX27811	Rhyolite	archeological	I Casteddi, Tavera	TAV19 H6 US301 Rhyo Percante	yes	no	no	yes
BDX27812	Obsidian	archeological	I Casteddi, Tavera	TAV19 HS penteW lamelle obsidienne	no	no	no	yes
R37_1	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_2	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_3	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_4	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_5	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_6	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_7	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_8	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_9	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no
R37_10	Flint	experimental	Unknown	Levallois debitage	yes	no	no	no

3 Photogrammetry Workflow

This section presents the workflow used for creating 3D models of the artefacts. Multiple photogrammetric protocols were employed. These protocols were tested both to explore different acquisition parameters and because the AUTOMATA protocol, optimised for robotic acquisition, is not sufficiently fast when applied manually. Some more traditional acquisition methods were therefore used to ensure an adequate number of 3D models could be produced.
The variety of the selected samples described in the previous section enabled thorough testing of the photogrammetric acquisition methodology, in accordance with what was already designed during T2.5. Challenges included the thinness of blades and bladelets, the transparency of unpatinated flint flakes, the reflectivity of highly polished materials such as obsidian and sigillata, and the extremely small size of some pieces, approaching the lower size limits defined in AUTOMATA (approximately 1 cm).

3.1 Description of the acquisition process

3.1.1 Acquisition process at INRIA

This section presents a series of acquisition tests conducted at INRIA with the device known as La Coupole, by Clément Joubert and Romain Pacanowski.

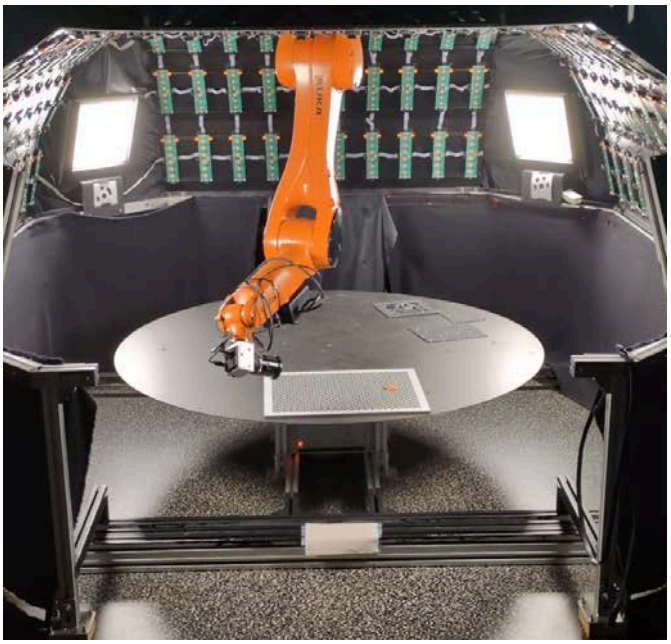

Fig. 2 - Picture of La Coupole acquiring a ceramic fragment.

The tests consist of taking multiple photos of each of the 20 selected ceramic artefacts and reconstructing their shapes and colours using RealityScan software (Epic Games, Inc., 2025). The device, La Coupole (fig. 2), is composed of:

- a Kuka robotic arm KR 10 R1100 sixx;

- a JAI SP 20000 Colour camera;
- a VS-Technology lens VS-L3528LM/F 35 mm;
- 4 LED panels for static illumination;
- a charuco board to calibrate the camera and support photogrammetry.

To digitise the artefacts, the camera was positioned and oriented around them using a dome pattern (fig. 3). The zenithal and azimuthal intervals were set to 20°, with an additional set of camera positions at a 70° zenith angle. Both sides of the object were acquired using 65 images per side. It took about 10 minutes to acquire one side of the object, meaning the robot took about 10 seconds to move and take a picture. This acquisition protocol is cautious, and the process should be faster in the final system, as the approximate bounding box of the object is expected to be available before the start of the acquisitions.
The picture resolution is 5120x3840 pixels; coupled with the lens, the achieved spatial resolution is ~77µm per pixel when the camera is 55 cm from the object. Pictures are saved in RAW format (.dng), and each picture is approximately 35 MB.

Fig. 3 - Distribution of the camera positions as a dome for photogrammetry.

Since the artefacts are thin, merging the object's different sides into a single mesh is quite difficult. To obtain our mesh (fig. 5), we followed the workflow described below.

- Reconstructing the upper side (~2-3 minutes) with RealityScan (Epic Games, Inc., 2025).
- Cleaning the output 3D mesh to obtain the masks of the object for each image. An example of the obtained masks is shown in fig. 4.
- Repeat the operation for the bottom side.
- Using images of both sides with their associated masks, reconstruct the full object.
- If the previous step does not work (which might happen because the sides are reconstructed independently), choose control points on the object and set their positions in each image (3 control points are generally enough). This part is currently very time-consuming because it is not automated.

Fig. 4 - Picture of a ceramic sample (left) with its associated masks (right).

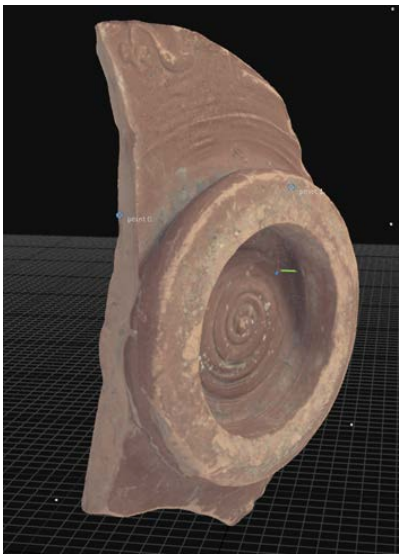

Fig. 5 - Textured 3D model of a ceramic sample in RealityScan software (Epic Games, Inc., 2025).

3.1.2 Acquisition process at Archeovision (UBM)

Here are the results of photogrammetric acquisition tests carried out at Archeovision (UBM). The tests were conducted by members of the Archeovision team, Mikaël Rouca and Pascal Mora, and are presented here in accordance with the adopted methodological approaches. The tests were conducted using different acquisition setups, with progressive refinement of supports, coded targets, and image-capture strategies.

3.1.2.1 Close-range photogrammetry

Test 01: sherd of terra sigillata (lot1n1)

Object to be digitised: ceramic object with an edge thickness of about 1 cm.

Photogrammetric issues:

- The thin edge limits the matching possibilities between the two faces.
- No expected difficulties related to the material.
- Limited depth of field.

Setup description:

- Object placed on a plexiglass plate.
- Printed target grid on the plate.
- Top and bottom photographs taken handheld.
- Ring flash lighting.

Results:
- Failure of matching between top and bottom images;
- Difficulty in taking bottom images handheld, especially at very low angles relative to the plate.
- A frame that is too small can occlude the object to be digitised.

Tests 02 and 03: sherds of terra sigillata (lot1n2 & lot1n3)

Object to be digitised: ceramic object with an edge thickness of about 1 cm.

Photogrammetric issues:
- The thin edge limits the possibilities for matching between the two faces.
- Limited depth of field.
- No expected difficulties related to the material.

Setup description:
- Object placed on a plexiglass plate-
- Four coded targets placed on the plexiglass.
- Top and bottom photographs taken handheld.
- Ring flash lighting.

Results:
- In test 02, failure of matching between top and bottom images.
- In test 03, the two faces could be aligned.
- Good target detection, but 4 targets are insufficient to ensure sufficient targets per image.
- Targets visible at very low angles are not detected.
- Difficulty in taking bottom images handheld, especially at low angles.
- Targets placed too close to the object can occlude it (when viewed from below), especially at low angles.

Tests 07, 08, and 09: lithic artefacts (BDX27802, BDX27804 & BDX27808)

Object to be digitised: sharp stone tools, approximately 2 to 7 cm in size.

Photogrammetric issues:
- The thin edge limits the matching possibilities between the two faces.
- Limited depth of field.

Setup description:
- Object placed on a plexiglass plate-

- Four double-sided coded targets placed on the plexiglass (front and back correspond to the same marker)-
- Photographs taken using a tripod-
- Ring flash lighting.

Results:

- The two faces are aligned thanks to the detected targets.
- The lower face (seen through the plexiglass) is not always reconstructed;
- Targets in images taken at very low angles are not detected, and these images do not align with the rest.
- The plexiglass scratches easily, and its reflections are problematic.

Possible future tests:

- Test the use of inclined targets in addition to horizontal targets to allow alignment of low-angle views;
- Test the use of anti-reflective glass.

Fig. 6 - Acquisition tests performed with a transparent tray (on the left) and recognition targets or on a solid rotating platform (on the right).

3.1.2.2 Resolution tests using focus stacking

Tests 10: lithic artefact (BDX27811)

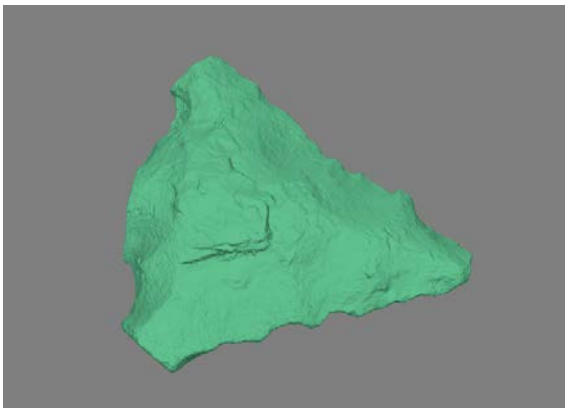

Fig. 7 - Digitisation of BDX27811 during test 10

Object to be digitised: flint object, approximately 2 cm.

Acquisition and processing details:

- Around 16 photos per stack;
- 99 stacks (12 to 20 photos per stack, depending on object orientation), produced by the camera;
- Several object positions;
- Use of masks during processing.

Processing times:

- Alignment: 1 minute;
- Model computation at maximum resolution: 6 minutes (5 million triangles).

Test 11: lithic artefact (BDX27810)

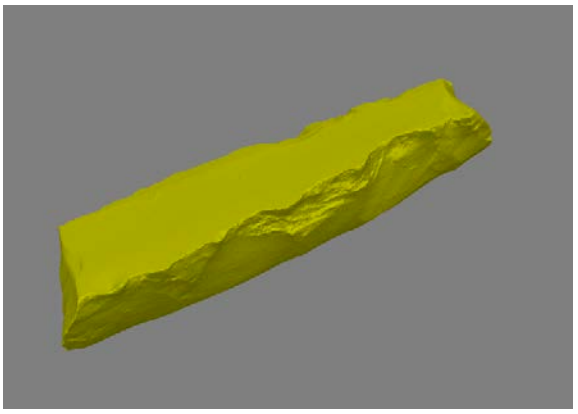

Fig. 8 - Digitisation of BDX27810 during test 11

Object to be digitised: chert, approximately 3 cm.

Acquisition and processing details:

- 56 stacks (12 to 20 photos per stack, depending on object orientation), produced by the camera;
- Several object positions;
- Use of masks during processing.

Processing times:

- Alignment: 1 minute;
- Model computation at maximum resolution: 8 minutes (7 million triangles).

Test 12: Ceramic artefact (not on the database yet)

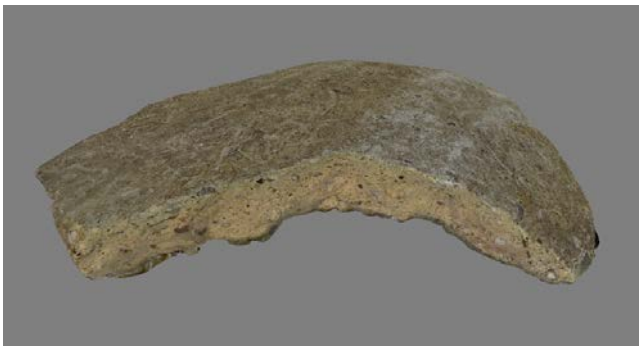

Fig. 9 - Digitisation of the "ceramic artefact" during test 12

Object size: approximately 5 cm.

Acquisition details:

- 71 handheld photographs;
- Use of masks and a plexiglass plate for the lower face.

Processing times:

- Alignment: 1 minute;
- Model computation at maximum resolution: 10 minutes (7 million triangles).

3.2 Role of AI in photogrammetry

Artificial intelligence plays a supporting role in AUTOMATA's photogrammetric workflow, primarily by improving efficiency, data reliability, and the suitability of 3D models for downstream archaeometric and robotic tasks. All AI-enabled operations are designed to run rapidly and, where possible, in real time, to support automated workflows required by AUTOMATA's enriched digitisation.

Given the high resolution of the primary 3D models produced during acquisition, often comprising millions of polygonal faces and reaching hundreds of megabytes per scan, AI-driven compression and simplification techniques are being investigated to generate **smart low-resolution representations** that significantly reduce computational demands. These methods identify regions

of high curvature or structural relevance and selectively reduce redundancy in low-information areas while preserving the geometric detail required for analyses sensitive to local shape variations. This optimisation step is essential for achieving near real-time performance, enabling the system to dynamically produce lightweight yet analytically reliable representations of the acquired objects.

AI-based **object segmentation** is another key component of the workflow. Before reconstruction, acquired images can be automatically masked to remove background elements, improving the cleanliness and reliability of the resulting 3D models. Different segmentation strategies were evaluated.

- OpenCV-based segmentation, relying on traditional computer-vision techniques such as colour thresholding, edge detection, background subtraction, or morphological filtering. Although lightweight and fast, its performance is sensitive to illumination changes and requires parameter tuning for different acquisition setups.
- The Segment Anything Model (SAM) provides a foundation model approach to segmentation and can quickly produce high-quality masks with minimal input. It is particularly effective when object boundaries are clear, but requires refinement for complex backgrounds.
- BiRefNet, a state-of-the-art deep neural architecture, specialises in extracting fine-grained foreground details. It uses bi-directional refinement stages to progressively sharpen object boundaries.
- LLM-based segmentation using Gemini exploits multimodal reasoning capabilities. Instead of relying solely on pixel-level cues, Gemini can interpret the scene contextually to guide segmentation, enabling flexible masking.

These four approaches are expected to be used in combination to generate more consistent and reliable foreground masks as the database grows and a wider range of acquisition conditions is captured. This step supports cleaner and more reliable model generation, reducing manual intervention and facilitating automated integration with other sensors.

Another important objective is using AI to **determine whether the collected photogrammetry data were acquired properly**. This includes detecting issues such as insufficient surface coverage, inconsistent viewpoints, or other acquisition anomalies that could affect the stability of the reconstruction.

AI-supported pre-processing of images and meshes is designed to operate in near real time, acting directly on the 3D model data as they are acquired. To meet the constraints of real deployment scenarios (e.g., operating speed and data stream volumes), pre-processing modules typically leverage lightweight neural architectures or machine learning algorithms that minimise latency while maintaining acceptable accuracy for the tasks to be executed.

The database stores the primary, unmodified acquisition data, but, for reasons of storage capacity, intermediate transformations, such as those discussed above, are not retained beyond the process for which they are necessary. This way, the architecture ensures more efficient data management and reduces I/O congestion.

4 Archaeometric Data Workflow

To test and validate the archaeometric data workflow, the sigillata fragments and lithic artefacts used for the 3D modelling procedures described in the previous sections were selected. Working with the same material ensures that appearance and compositional data can later be aligned, compared, and integrated into a coherent dataset.

At this stage, the archaeometric analyses have been carried out manually, following standard procedures for HSI, pXRF and Raman spectrometry. Although manual processing provides reliable reference datasets, the heterogeneity across partners' workflows underscores the need for a unified approach. To process these datasets effectively, a common workflow is therefore required. Existing analytical data vary in format, structure and level of preprocessing, as they were produced independently. Establishing a shared framework is essential to harmonise these datasets and ensure they are suitable for subsequent processing steps.

The proposed methodology involves defining consistent protocols for data normalisation and for integrating outputs from the various analytical techniques. As a practical solution for designing and testing these procedures, it was decided to use Orange (Demšar et al., 2013), an open-source visual data-analysis platform. Orange supports the construction of modular workflows capable of combining inputs from HSI, pXRF and Raman spectroscopy within a single environment.

A key advantage of this approach is that the visual workflows developed in Orange can be used as Python scripts, enabling their incorporation into broader automated systems as the project progresses. This aligns with the decision to develop all AI and modelling tools in Python, ensuring interoperability, consistency and long-term maintainability across the entire AUTOMATA framework.

4.1 Hyperspectral Imaging

Following the 3D digitisation, hyperspectral imaging (HSI) is planned as a standard component of the AUTOMATA working pipeline.

During the current testing phase, HSI analyses were conducted on the selected ceramic and lithic fragments using a Specim IQ camera. These acquisitions were performed under different controlled settings within the Archéosciences laboratory at UBM and the LightTECH photonics and laser microstructuring lab in the University of Bordeaux (figs. 10-11), employing the standard operating procedures defined for the instrument (see Deliverable 2.1, section 1.4.1.1, for further details on the camera's operating principles and configuration requirements).

For each ceramic and lithic fragment, a dedicated hyperspectral image was captured to document its spectral characteristics and serve as a basis for subsequent processing and integration.

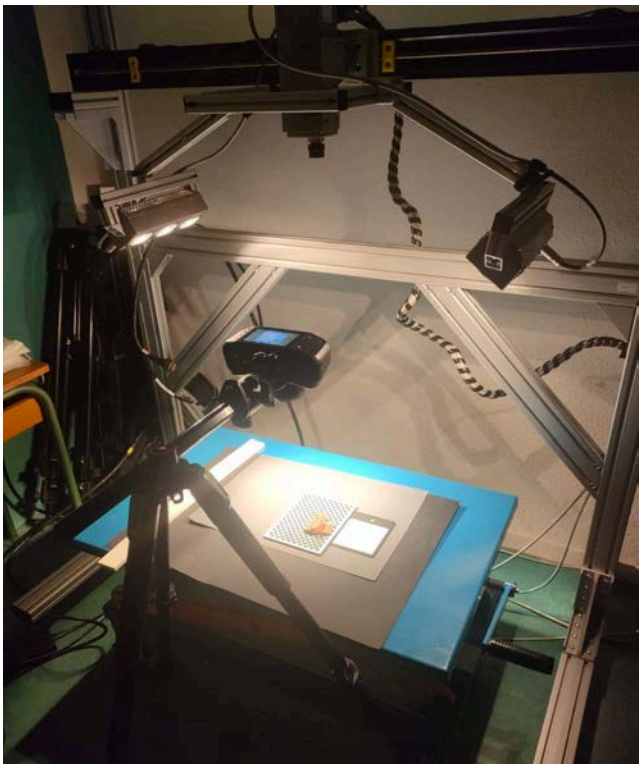

Fig. 10 - Specim IQ camera mounted on a tripod at the Archéosciences laboratory (UBM). For these acquisitions, the objects are illuminated by two bars of three halogen lamps, oriented at 45°.

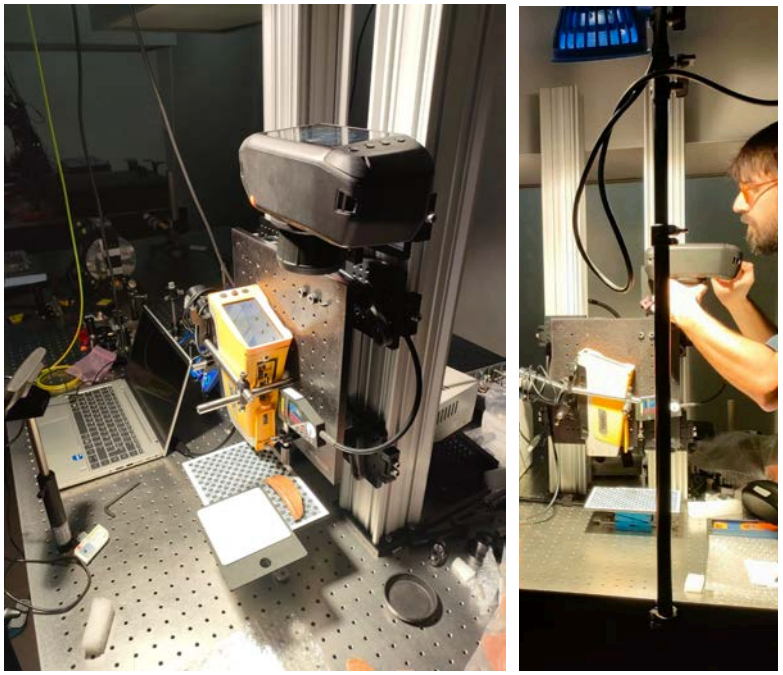

Fig. 11 - Specim IQ camera at the LightTECH photonics and laser microstructuring lab.

Every measurement results in a comprehensive collection of raw sensor data, calibration frames and derived imagery, documenting the complete spectral content of the fragment. The raw output includes sensor data, calibration frames, and the associated metadata required to reconstruct the reflectance cube. In addition, the camera automatically generates several preview images for rapid visual inspection. The single acquisition generates the following groups of files:

1. **Capture files (raw sensor data)**
   These files contain the unprocessed hyperspectral information as recorded by the sensor:
   - sample.raw – the full raw datacube for the fragment (one value per pixel per wavelength band).
   - sample.hdr – header file describing the structure of the raw datacube (image dimensions, number of bands, wavelengths, bit depth).
   - DARKREF_sample.raw / DARKREF_sample.hdr – dark-reference frames used to remove sensor noise.
   - WHITEREF_sample.raw / WHITEREF_sample.hdr – white-reference frames used for radiometric calibration.

   These files are essential for converting the raw sensor output into reflectance values.

2. **Processed reflectance data**
   After calibration, the camera creates the reflectance datacube:
   - REFLECTANCE_sample.dat – the full reflectance cube (very large file), containing calibrated spectral values for each pixel across all wavelengths.
   - REFLECTANCE_sample.hdr – header describing the reflectance cube.
   - REFLECTANCE_sample.png – a visual rendering of the reflectance data for quick inspection.

     These files form the core dataset for subsequent processing and advanced analysis.

3. **Preview and derived images**
   To support quick visual evaluation, the system produces conventional RGB outputs:
   - RGBSCENE_sample.png – RGB rendering of the scene.
   - RGBVIEWFINDER_sample.png – the image as captured through the camera's viewfinder.
   - RGBBACKGROUND_sample.png – background RGB image used during the acquisition process.
   - sample.png, spettrosample.png – additional previews and spectral plots automatically generated.

4. **Metadata files**
   These record information about the acquisition parameters:
   - manifest.xml – global metadata for the acquisition session.
   - metadata/sample.xml – detailed metadata for the specific capture (exposure, integration time, illumination, camera settings).
   - .validated – internal file marking the dataset as correctly saved.

Once the reflectance data have been generated, the next step of the workflow involves processing the hyperspectral cube using the Orange data-analysis platform (fig. 12). The **reflectance.hdr** file is first loaded into the software, which allows the corresponding datacube to be visualised and manipulated. The initial steps involve applying a white-reference correction and isolating the artefact by masking the background and shadows, ensuring that subsequent analyses focus solely on the ceramic or lithic surface. Moreover, this step reduces the number of variables that will be processed.

Principal Component Analysis (PCA) is then applied to the datacube to highlight spectral variability and reveal compositional or surface features that may not be visible in conventional RGB images. PCA allows the major sources of variance across the spectral bands to be explored, often highlighting differences in fabrics, inclusions, surface treatments or areas affected by alteration.

Independent Component Analysis (ICA) can be employed, as an alternative or complementary method to PCA, to further interrogate the hyperspectral dataset. ICA seeks to decompose the data into statistically independent sources by exploiting higher-order statistical structure and deviations from Gaussianity. It is particularly effective at separating mixed pixel signatures into latent spectral components that may correspond to distinct materials, surface treatments, or alteration products. Consequently, PCA provides a robust initial reduction and exploration of variance within the dataset, while ICA serves as a complementary method aimed at isolating discrete spectral sources and enhancing the detection of subtle or obscured compositional features. Both analyses can be applied to explore the datacube and reduce the number of variables.

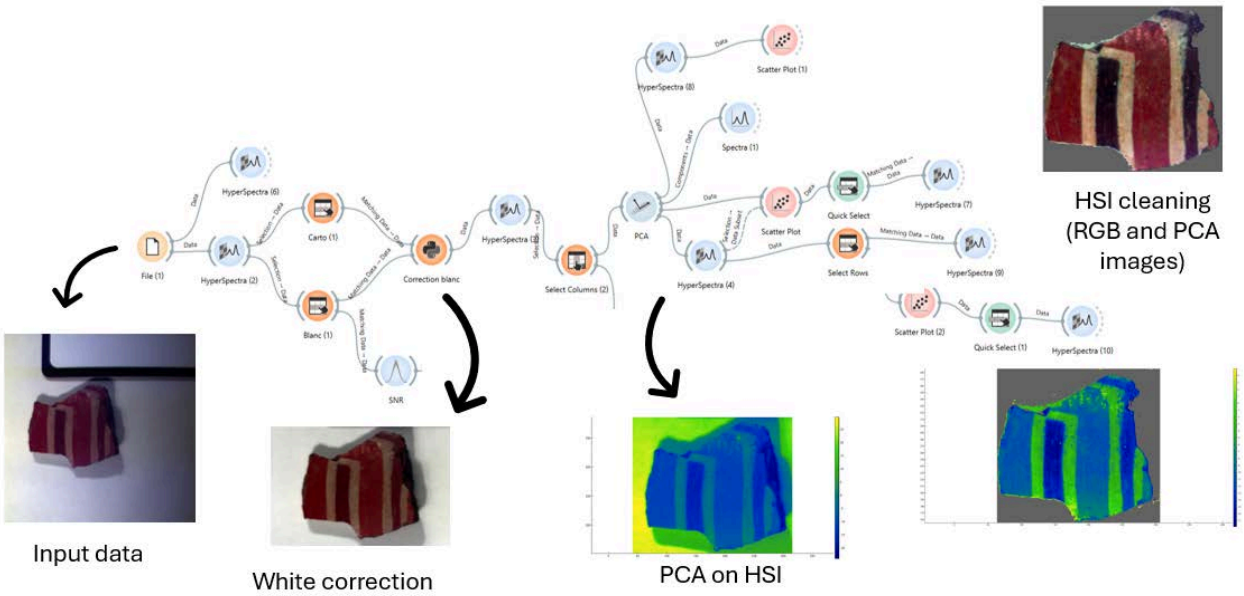

Fig. 12 - Overview of the initial HSI processing workflow implemented with Orange, showing input acquisition, white correction, PCA computation, and subsequent cleaning of RGB and PCA-derived images.

Following the PCA, individual spectra can be extracted from specific regions of interest on the artefact. This allows the comparison of spectral signatures across different surfaces, textures or

manufacturing features. The software provides interactive tools for selecting pixels or areas directly on the reflectance or PCA-rendered images.

The processing workflow typically generates:

- **Derived images**, including PCA component images (e.g., PC1, PC2, PC3) or false-colour composites based on selected principal components;
- **Spectral plots** for each selected pixel or region of interest, showing reflectance values across the full wavelength range;
- **Tabular outputs** (e.g., .csv files) containing numerical spectral data for further statistical treatment;
- **Processed datacubes** or masks that can be exported for later integration into the AUTOMATA database or for use in machine-learning pipelines;
- **Visual previews** (PNG images) summarising the PCA results, masks, or extracted features.

These outputs serve as the basis for validating the archaeometric workflow and for future integration into automated pipelines using Python-based tools.

4.2 Portable XRF

Following the workflow outlined in Deliverable 2.1, the results of 3D modelling and hyperspectral imaging (HSI) inform the decision on whether to carry out additional analyses. This depends on the artefact's surface characteristics and the availability of sufficiently large, flat areas suitable for accurate sensor readings. When these conditions are met, portable X-ray fluorescence (pXRF) may be used (see Deliverable 2.1, paragraph 1.4.1.2) to investigate the composition of ceramic pastes or lithic materials. The selection of measurement points is informed by the regions of interest identified through HSI, based on variations in spectral response and surface morphology.
During this testing phase, following the same approach described for HSI in the previous section, pXRF measurements were performed on the subset of twenty sigillata ceramic fragments and selected lithic specimens presented in Section 2 of this deliverable. Analyses were conducted using an Olympus VANTA C-series handheld X-ray fluorescence analyser in two laboratory settings, mirroring the HSI acquisition workflow: the Archéosciences laboratory at Université Bordeaux Montaigne and the LightTECH photonics and laser micro-structuring laboratory in Bordeaux (fig. 13). All measurements followed the instrument's standard operating procedures (see Deliverable 2.1). For each sample, between 1 and 3 measurements were taken, with only a single acquisition when the fragment's physical characteristics did not allow more.

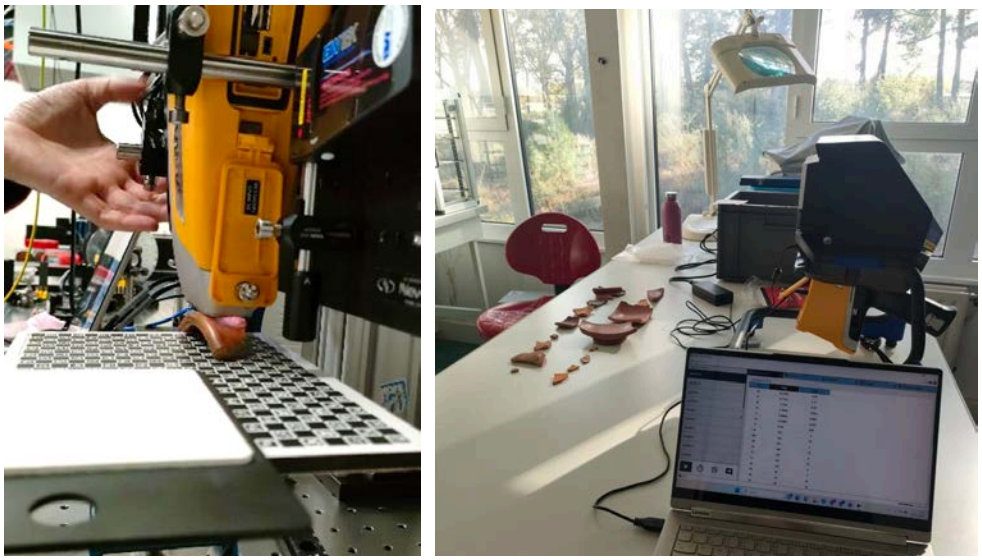

Fig. 13 - Left: Olympus Vanta pXRF mounted on a stand at the LightTECH photonics and laser microstructuring laboratory. Right: Olympus Vanta pXRF on its field stand at the Archéosciences laboratory (UBM).

Each pXRF acquisition session generates a structured set of output files compiled by the instrument software into a batch. For the Olympus VANTA system used in the testing phase, these consist of 3 files, as described below.

- Beamspectra files (**beamspectra-*.csv**), containing the full X-ray emission spectra for all measurements in the batch. These CSV files record the raw spectral counts and associated parameters for each acquisition point.
- Chemistry files (**chemistry-*.csv**), summarising the processed elemental composition results for every measurement performed during the session. Each row corresponds to an individual analysis point, including element concentrations and statistical parameters.
- An accompanying image directory (**chemistry-*-images/**), which stores automatically generated images for each measurement.

Before the pXRF results can be explored within the AUTOMATA software and stored in the project database, the data produced by the instrument need to be prepared and reorganised. This is because the raw output files generated during each acquisition session are primarily designed for immediate inspection within the instrument's proprietary environment and therefore require adaptation to support long-term storage, interoperability, and cross-linking with other datasets.
First, as AUTOMATA treats each pXRF measurement as an independent analytical record, the batch-based outputs must be converted into a set of individual entries. This step ensures that each spectrum can be unambiguously associated with a single artefact, with the specific measurement location identified through HSI and photogrammetry, and with the relevant acquisition metadata.
Then, since each individual measurement is recorded using two different X-ray beam settings, designed to target different ranges of chemical elements, the data from the two beams in the beamspectra-*.csv needs to be combined into one representative spectrum, integrating complementary information on light and heavier elements to produce a single analytical output per measurement point.

Also, the chemistry.csv original spreadsheet produced by the instrument includes acquisition metadata, multiple result types, and elemental values that are not all relevant for further analysis. For this reason, only the concentration data for measured elements are retained, while auxiliary columns and values below the limit of detection are removed. Each measurement is clearly labelled to ensure traceability, and calibration checks and reference standards, typically recorded at the beginning and at the end of each session, are identified and treated separately from archaeological samples.
The cleaned chemistry-*.csv and beamspectra-*.csv files can then be imported into the Orange data mining environment for processing and exploration. Chemistry data are used to investigate quantitative elemental or oxide compositions through descriptive statistics, data normalisation, and multivariate analyses such as principal component analysis and clustering, with the resulting quantitative outputs exported as structured **CSV files**. In parallel, the processed beamspectra data allow spectral profiles to be inspected and compared, supporting the identification of patterns, similarities, and potential outliers across measurements. Spectral data can also be visualised and saved as **images** (.png or .pdf), providing an immediate graphical representation of the X-ray emission characteristics associated with each acquisition.
While these processed outputs are not directly required in the RIS3D environment, they are essential for structuring and documenting pXRF results in the AUTOMATA database, ensuring consistency, reproducibility, and reliable linkage to the corresponding 3D models and imaging datasets.

4.3 Raman Spectrometry

The final analytical technique considered within the workflow, although not applied in all cases, is Raman spectrometry, as outlined in Deliverable 2.1 (Section 3). During the testing phase, and following the same methodological approach adopted for HSI and pXRF in the previous sections, Raman measurements were carried out on the subset of twenty sigillata ceramic fragments and selected lithic specimens presented in Section 2 of this deliverable.
For the testing activities, an i-Raman Plus 785H spectrometer (Metrohm–BWTek) was employed. The instrument is available at the Archéosciences laboratory in Bordeaux and was used in two different laboratory settings, mirroring the acquisition workflow established for HSI and pXRF: the Archéosciences laboratory at Université Bordeaux Montaigne and the LightTECH photonics and laser micro-structuring laboratory in Bordeaux (fig. 14).
All Raman analyses were performed in accordance with the standard operating procedures defined for the instrument (see Deliverable 2.1). For each sample, between one and three measurements were acquired, with a single measurement carried out when the physical characteristics or surface condition of the fragment did not permit multiple acquisitions.

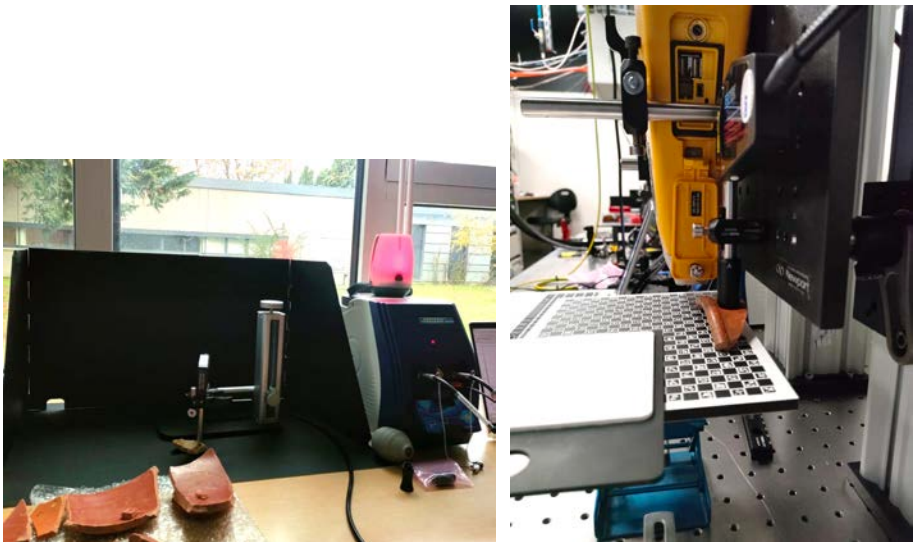

Fig. 14 - Left: i-Raman Plus 785H Raman spectrometer (B&W Tek, Metrohm) used for the analysis of ceramic fragments at the Archéosciences laboratory (UBM). Right: the same spectrometer installed at the LightTECH Photonics and Laser Microstructuring Laboratory.

All analysis results were recorded in a **.txtr** file. This file can be easily read by the Orange software, in particular through the Oranchada add-on (Georgiev et al., 2025) (fig. 15).

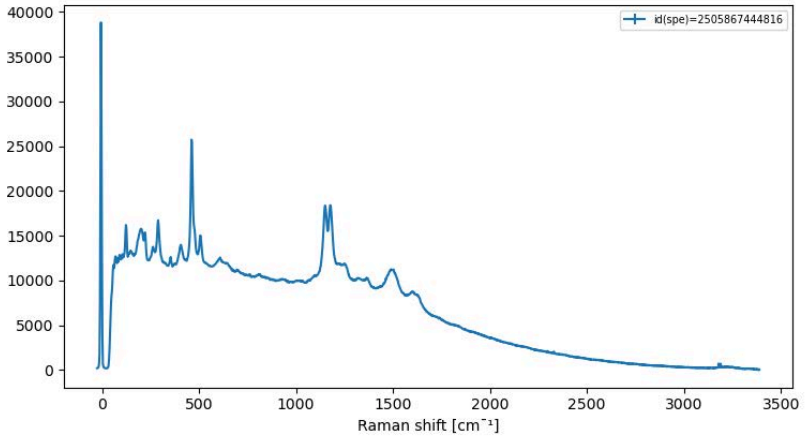

Fig. 15 - Raman spectra from BDX27801 generated by Oranchada.

4.4 Role of AI in the archaeometric workflow

The automation of archaeometric data acquisition and processing will be progressively achieved through the integration of AI and ML algorithms. Current experimentation focuses on selected stages of the pipeline where AI/ML can increase processing speed and efficiency, and enhance data consistency and processing robustness. At present, such methods are being explored and implemented in accordance with the steps described below.

- **Identification of archaeologically and morphologically significant areas.** By combining information derived from 3D geometry and HSI, AI models are being trained to detect areas

of archaeological interest on artefact surfaces. Colourimetric and spectral features are used as primary descriptors to support the recognition of suitable shapes for further analyses, manufacturing traces, and zones of alteration.

- **Detection of noisy or unreliable analytical signals.** ML algorithms are applied to identify anomalous or noisy signals in spectroscopic datasets (pXRF and Raman), enabling the detection of incorrect acquisition parameters or instrument miscalibration.

At this stage, AI integration remains experimental and complementary to manual procedures. Full and effective workflow automation can only be achieved once the current archaeometric pipeline has been fully defined and standardised.

5 Data Management and Database Infrastructure

The AUTOMATA database infrastructure consolidates all photogrammetry and archaeometric datasets described in the preceding sections, ensuring that each digital output produced during acquisition and processing is securely stored, consistently referenced, and readily interoperable. All working data are maintained within an **ArcheoGRID sandbox**, which functions as the project's internal and controlled environment for storage, indexing, and data management throughout the development phase.

Datasets produced by partners during digitisation are transferred to the ArcheoGRID sandbox via secure SFTP. Once ingested, each dataset is assigned a stable internal identifier and linked to its associated object, context, and analytical record. This guarantees coherent relationships between geometric, visual, and physico-chemical layers, allowing the database to serve as a unified reference point for all materials documented in AUTOMATA.

The ArcheoGRID environment supports the controlled handling of both raw and processed files, as well as the structured representation of metadata generated across the photogrammetry and archaeometric workflows, organised within a hierarchical directory structure that preserves provenance and supports reproducibility (fig. 16).

For the several types of data produced within AUTOMATA, users may wish to work with simplified/compressed versions when performing downstream analyses or visualisation tasks. Such versions are typically easier to handle, faster to process, and more compatible with a wider range of external software tools. At the same time, the original acquisition outputs are retained as the authoritative source data within the database. To avoid unnecessary duplication, the AUTOMATA workflow stores both the original and simplified files, and does not archive simplified versions as separate data objects. Instead, the tools for generating simplified representations are provided as part of the workflow and accessible through the graphical user interface (GUI). This approach allows users to decide, according to their specific needs, whether to work with the original data or generate a simplified representation on demand, while ensuring that all derived products remain reproducible from the stored source files.

Metadata from the acquisition and processing stages are stored within a PostgreSQL relational database, complemented by JSON/JSONB fields for recording variable or technique-specific parameters. This hybrid structure accommodates heterogeneous archaeometric outputs while maintaining a consistent and queryable core schema across the project.

A single .xml file is associated with each analysis batch. It contains all metadata for the batch, including that for all analysed artefacts and all technologies used during that batch. Using a single file greatly facilitates metadata use and reduces the risk of losing the link between source files and their metadata. However, this structure is designed for automation and is currently not very convenient for manual data entry. Spatial metadata follow a relative coordinate system used during acquisition and are later aligned to the corresponding 3D model. This ensures interoperability between datasets and enables precise cross-referencing of analytical points (such as hyperspectral imaging areas, pXRF spots or Raman measurements) within later integration environments.

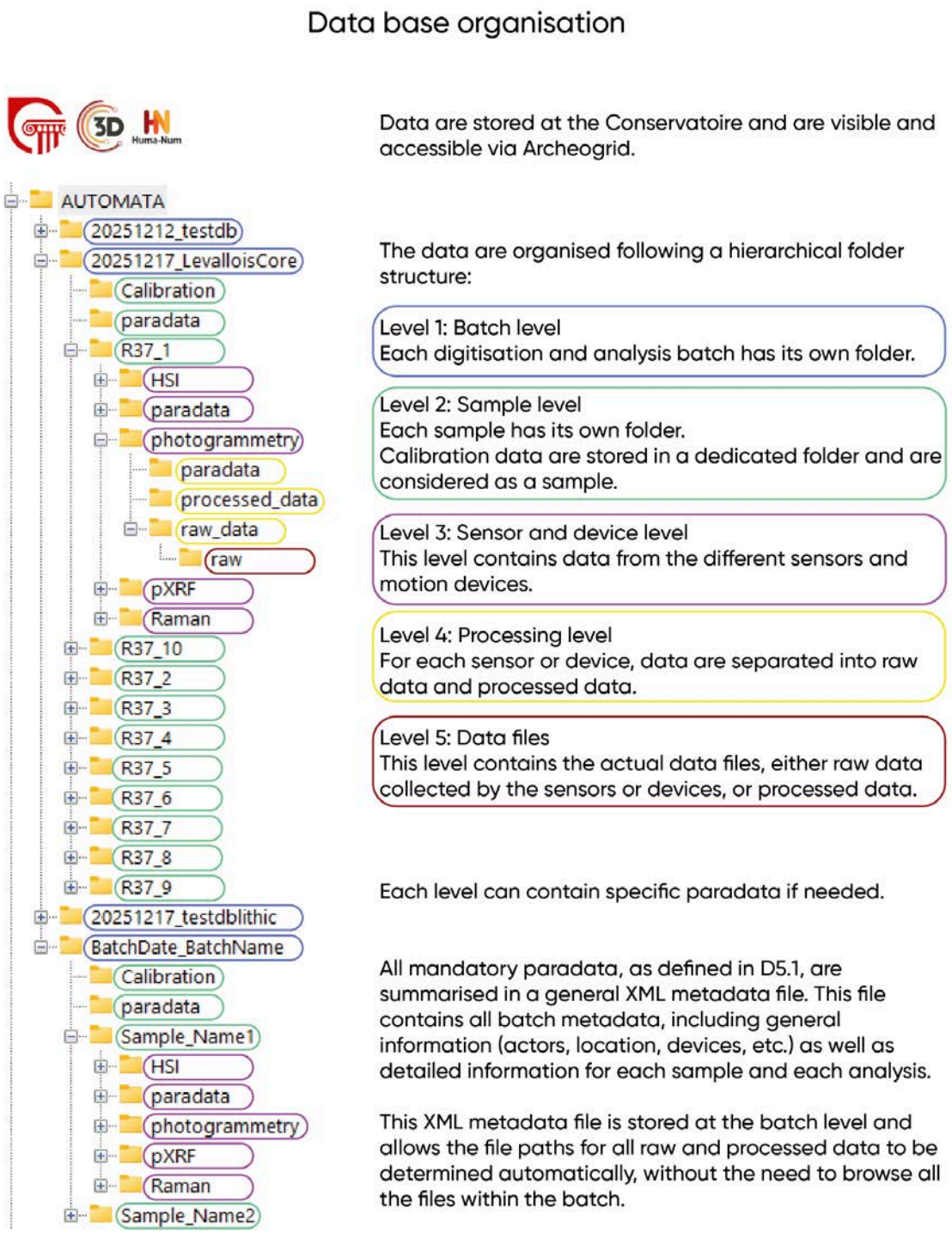

Fig. 16 - Example of the hierarchical directory structure implemented within the ArcheoGRID sandbox for the AUTOMATA project, from batch-level and object-level folders to technique-specific raw data, processed data, and associated paradata.

The internal GUI allows project partners to access, browse, and query datasets stored in the ArcheoGrid database. Through this interface, users can view object-level records, explore linked analytical datasets, and verify metadata consistency. The GUI draws directly from the PostgreSQL–JSON infrastructure, supporting both structured querying and flexible metadata navigation tied to the workflows described in earlier sections.

The database serves as the foundation for the next stage of development: the integration of AUTOMATA datasets into the RIS3D environment. This step, which will be detailed in Deliverable 5.3 (*Reference enriched 3D data*), will enable the combined visualisation of 3D models, spatial metadata, and analytical data within a unified interface. At this stage, RIS3D is referenced only as the forthcoming integration platform; the present deliverable focuses exclusively on the creation and structure of the database itself.
During the project, partners maintain local physical backups to ensure data security at all acquisition stages. Once datasets reach their final validated form, they will be deposited with the Archaeology Data Service (ADS) for long-term preservation, dissemination, and FAIR-compliant accessibility. ADS supports all file formats used in AUTOMATA and provides persistent identifiers, certified archival standards, and broad discoverability through platforms such as ArchSearch, ARIADNE, and Europeana.

Link to the database: https://www-dev.archeogrid.fr/project/12527

Link to the XML metadata for Levallois testing data set, which will be our basis for all metadata files: https://www-dev.archeogrid.fr/viewer/12697_137584?format=hhdd

Tab. 2. Overview of AUTOMATA acquisition outputs

Data stream	Raw acquisition outputs (authoritative source data)	Processed / derived outputs (examples)	Metadata / paradata files	Storage & linkage logic
Photogrammetry data (RealityScan / tests)	RAW images: **.dng, .nef**	Textured 3D model and intermediate products (e.g., masks, stacked images)	Acquisition parameters recorded as part of the dataset; linked to object record	Stored in the ArcheoGRID sandbox within the object-level folder; raw images retained as source data, with simplified products generated on demand to avoid duplications. Linked to object identifier and subsequent RIS3D ingestion.

Archaeometric data – Hyperspectral imaging (HSI, Specim IQ)	Capture (raw sensor data): **sample.raw**, **sample.hdr**, **DARKREF_sample.raw/.hdr**, **WHITEREF_sample.raw/.hdr**	Reflectance cube: **REFLECTANCE_sample.dat**, **REFLECTANCE_sample.hdr**, quicklook **REFLECTANCE_sample.png**; RGB previews (**RGBSCENE_sample.png**, **RGBVIEWFINDER_sample.png**, **RGBBACKGROUND_sample.png**, sample.png). Derived images, spectral plots, tabular outputs (e.g. csv files), processed datacubes.	**manifest.xml**; **metadata/sample.xml**; **.validated**	Raw + calibrated products stored in technique-specific folders under the object dataset; metadata files retained alongside the acquisition set to preserve provenance and reproducibility; prepared to feed RIS3D anchoring/alignment steps.
Archaeometric data – Portable XRF (pXRF, Olympus VANTA)	Batch outputs: **beamspectra-*.csv**, **chemistry-*.csv**, **chemistry-*-images/**	Cleaned / reorganised CSVs for analysis; optional plot exports **.png / .pdf**	Batch-level metadata referenced through the single XML strategy (see below)	Batch outputs are ingested, then reorganised so each measurement can be treated as an independent analytical record and linked to the correct object + measurement context; outputs remain reproducible from stored source files.

Archaeometric data – Raman spectrometry (i-Raman Plus 785H)	Spectra recorded as **.txtr**	Derived visualisations via Orange/Orange hada (as needed)	Batch-level metadata referenced through the single XML strategy (see below)	Raman outputs stored under technique-specific folders; linked to object record and later spatial anchoring workflow (RIS3D) where applicable.
Cross-technique metadata layer (hybrid "3D database" backbone)	N/A (metadata layer)	N/A	**One .xml file per analysis batch** (contains metadata for all artefacts + technologies within the batch); plus PostgreSQL fields (relational core + JSON/JSONB for technique-specific parameters)	This is the "glue" between file-based assets and queryable records: stable internal identifiers, provenance/traceability, and preparation for RIS3D integration (object-level, spatially anchored querying later on).
Downstream analysis outputs (current testing; future automation)	N/A (derived from stored source data)	HSI: PCA/ICA images, false-colour composites, extracted spectra, **.csv** tables; pXRF: processed tables/plots, exported structured outputs	Stored as processed/derived products when needed; always reproducible from source data	Keep derived products clearly separated from authoritative source data; document software/toolchain used so results are reproducible and traceable.

General rules (applied to all data types):

- Raw and processed assets are stored in the structured repository, while descriptive, provenance and parameter information is stored in the metadata layer (PostgreSQL with JSON/JSONB extensions), with batch XML used where relevant for campaign-level ingestion.
- Each asset is registered with a persistent internal identifier and linked to the relevant artefact/context and acquisition/analysis event records.

- Derivatives intended for visualisation or efficiency (e.g., simplified models) are generated reproducibly from authoritative sources; when stored, they are explicitly linked to their parent assets and processing parameters.
- At minimum, ingestion includes checksum computation and a validation status (uploaded / validated / curated) to support auditability across partners.
- Spatial coordinates recorded during acquisition as relative references are preserved and aligned to the 3D mesh space in RIS3D (D5.3) to enable spatially anchored enrichment.

5.1 Storage

File size is a key operational constraint for database ingestion. Uploading large and/or numerous assets can quickly become a bottleneck, so storage management is treated as an integral part of the acquisition workflow, not as a post-hoc activity.

In line with the principles set out in Deliverable 10.1 (Data Management Plan), we therefore prioritise the retention of authoritative source data and essential processed outputs, while minimising unnecessary duplication and avoiding the systematic preservation of intermediate files that can be reliably regenerated from documented workflows.

For hyperspectral files, the Orange workflow already supports automated normalisation using the white reference and the retention of only the image region containing the sample. This approach is used to reduce volume while preserving analytical integrity and reproducibility.

For 3D models, file size is primarily driven by model resolution. Resolution choices are treated as a controlled parameter and will be tuned against the actual operational needs of archaeometric acquisition and interpretation, so that models remain fit for purpose without generating avoidable storage overhead.

Photogrammetry image sets are the most storage-intensive component. To keep ingestion sustainable, the acquisition strategy is oriented towards collecting only the images required to reconstruct a reliable 3D model (with dedicated tests defining the minimum acquisition set for La Coupole datasets) and towards streamlining image retention in a way that remains consistent with controlled lighting and calibration conditions within AUTOMATA. Any conversion strategy for acquisition formats (including the handling of RAW versus JPEG) will be formalised within the project's data governance and documented accordingly, so that quality requirements, calibration assumptions, and provenance remain explicit and auditable.

Stacked imaging (used when the depth of field is insufficient) can substantially increase storage volume. For this reason, stacking is treated as an exception driven by specific acquisition constraints, and it is avoided whenever standard acquisition configurations can deliver adequate focus and reconstruction quality.

6 Conclusions and future development

The AUTOMATA database (data backbone) has been implemented on ArcheoGRID and is accessible online within the project environment. It is built upon the metadata scheme developed in Deliverable D5.1 (*Ontology and Metadata Scheme for Enriched Digitisation*). The current version represents an initial iteration that consolidates the core infrastructure (repository organisation, persistent identifiers, and metadata registration) required to manage the first complete acquisition campaigns.. The next implementation steps are expected during the implementation of tasks T5.3 (*Creation of the enriched 3D model*) and T5.4 (*Automation of enriched 3D models generation*), as well as throughout the entire duration of WP6 (*Technologies for visualisation and processing of enriched digitisation*). These next activities will operationalise the spatial anchoring of analytical measurements to 3D meshes (including vertex-level localisation where appropriate), building on the data structures and traceability mechanisms established in this deliverable.

The database has been designed to manage different levels of data processing, including raw data and processed datasets. This multi-level structure ensures that the data can support both immediate experimental needs and long-term reuse, while remaining compatible with future automation and visualisation tools. Importantly, the tests conducted on different objects and across multiple analytical techniques were used not only to verify feasibility, but also to derive practical requirements for the backbone: they informed which parameters must be captured systematically, which elements require technique-specific flexibility, and how raw assets, processed outputs, and provenance information should be linked for reliable reuse and downstream integration.

A single .xml file is associated with each analysis batch and contains all the metadata related to the batch, including metadata for all analysed artefacts and for all technologies used during that batch. This approach provides a coherent packaging unit for ingestion and traceability across heterogeneous datasets. Nevertheless, some parameters relevant to robotic acquisition and automated production are not yet represented; these will be incorporated through controlled, versioned updates to the metadata model, aligned with WP10 data governance and the implementation needs of T5.3–T5.4.

The initial sample set supported iterative testing—particularly for 3D digitisation—and enabled protocol refinement while the acquisition procedures were being stabilised. The resulting datasets therefore serve a dual purpose: they provide reference material for quality control and, at the same time, they anchor the definition of the data structures required to scale acquisition and ingestion across partners and technologies. At present, manually carrying out 3D digitisation, archaeometric analyses, metadata recording, and uploading files to the database is laborious and time-consuming. As a result, the first complete datasets will be uploaded to the database progressively, following validation and traceability checks. As automation components are introduced in T5.4, the manual burden associated with packaging, metadata registration, and ingestion will be reduced, improving throughput while preserving provenance and reproducibility.

References

Epic Games, Inc. (2025). *RealityScan* (Versione 2.0.1) [Software]. https://www.realityscan.com

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*, 2349–2353.

Georgiev, G., Coca-Lopez, N., Lellinger, D., Iliev, L., Marinov, E., Tsoneva, S., Kochev, N., Bañares, M. A., Portela, R., & Jeliazkova, N. (2025). Open source for Raman spectroscopy data harmonization. *Journal of Raman Spectroscopy, 56*(9), 878–881. https://doi.org/10.1002/jrs.6789