This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

GRANT AGREEMENT NUMBER:	**693548**
PROJECT ACRONYM:	**ArchAIDE**
PROJECT TITLE:	**Archaeological Automatic Interpretation and Documentatic of cEramics**
FUNDING SCHEME:	**H2020-REFLECTIVE-6-2015**
PROJECT COORDINATOR	**Prof Maria Letizia Gualandi, UNIPI**
TEL:	**+39 05022 15817**
E-MAIL:	**maria.letizia.gualandi@unipi.it**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

Doc Title. ORDP: Open Research Data Pilot

D N° 10.2

version: 0.2

Revision: Final Version

Work Package	10
Lead Author (Org)	Tim Evans (UoY)
Contributing Author(s) (Org)	
Due Date	M6
Date	30th November 2016

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

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
0.1	22.11.2016	Tim Evans	Complete Draft
0.2	25.11.2016	Tim Evans	Final version

Disclaimer

Table of Contents

Abbreviations

WP: Work package

M: Month

UNIPI: Università di Pisa

UoY: University of York

UB: Universitat de Barcelona

UCO: Universitaet zu Koeln

TAU: Tel Aviv University

CNR: Consiglio Nazionale delle Ricerche

INERA: Inera srl

BARAKA: Baraka Arqueologos S.L.

Elements: Elements centro de gestio i difusio de patrimoni cultural

Executive summary

The formal Data Management plan consists of an online document written in the templates within the 'DMPonline' tool: part of the the Open Research Data Pilot (ORD) funded under Horizon 2020. The ArchAIDE DMP is live at: https://dmponline.dcc.ac.uk/projects/archaide-horizon-2020-dmp

The online DMP is available to view at the above URL. According to the ORD guidance the DMP is to be considered as a living document, with edits implemented over the course of the ArchAIDE project. The document consists of three elements:

1. Initial DMP: a first version of the DMP to be submitted within the first six months of the project
2. Detailed DMP: updated over the course of the project whenever significant changes have arisen.
3. Final review DMP: reflecting all updates made over the course of the project.

The ArchAIDE European project aims at developing a highly innovative application for the archaeological practice, which can quickly recognize potsherds and improve dating and classification systems. The project, funded under the Horizon 2020 European programme, is coordinated by the researchers of the University of Pisa. ArchAIDE aims at improving access and promotion of the European archaeological heritage through the development and implementation of an open-data database, which will allow all application users to use this information. All research data collected and generated during the project will be managed securely during the project lifetime, made available as Open Access data by the project end, and securely preserved in the Archaeology Data Service (ADS) repository into perpetuity. This will include textual data and visual data (photographs, vector and raster images/drawing, eventually 3D models), which will be collected and documented according to the internationally agreed standards set out in the ADS/ Digital Antiquity Guides to Good Practice (http://guides.archaeologydataservice.ac.uk). Linked open data held in the ADS RDF triplestore will provide an alternative means of access to the data, via a SPARQL query endpoint.

The Project Data Contact is Tim Evans (Archaeology Data Service) tim.evans@york.ac.uk

1. Data summary

- The purpose of data collection is to populate a database that will act as automated reference tool for the recognition and classification of pottery sherds from archaeological excavations.
- The reference database - where copyright has been cleared - will be publicly available under the standard ADS Terms and Conditions of Use.
- The primary data type will be the database itself which will incorporate textual data, raster and vector images, and 3D models.
- The database will incorporate data from existing sources including the Roman Amphorae digital resource (http://dx.doi.org/10.5284/1028192)
- The dataset will provide a reference resource for archaeological ceramic specialists and non-specialists alike.

2. Fair Data

2.1. Making data findable, including provisions for metadata:

- The final dataset will be archived by the Archaeology Data Service (ADS) as a single collection. Collection-level metadata (based on Dublin Core) will be created, which will allow the resource to be found within the main ADS website. This metadata will also by exposed/consumed by other portals such as ARIADNE. In addition, it is also planned to publish the dataset as Linked Open Data via the stores within Allegrograph, and published via Pubby and the ADS' SPARQL interface.
- The ADS archive will be identifiable via a Digital Object Identifier (DOI), registered with Datacite.
- ADS Collection-level metadata is based on Dublin Core (DC) elements. DC.Subject terms are based on archaeology/heritage specific thesauri and vocabularies updated and maintained as Linked Open Data (LOD) by national cultural heritage bodies (see http://www.heritagedata.org/). These allow subject terms such as 'CERAMIC' to be meaningfully and consistently recorded. As part of the ongoing ARIADNE project these terms have also been mapped to the Ariadne Dataset Catalogue Model (ACDM see http://portal.ariadne-infrastructure.eu/about)
- Over the course of data collection a clear versioning system - aided by consistent file-naming strategy) will be used, based on the guidelines stipulated in the Archaeology Data Service / Digital Antiquity Guides to Good Practice.
- As outlined above, the final archive will reside with the ADS with metadata compiled to their standards, based on DC terms. Existing heritage thesauri will be used for the recording of subject terms

2.2. Making data openly accessible:

- The main output of the project will be the project reference database. This database will be archived with the Archaeology Data Service (ADS). This database - with the exception of material not copyright cleared - will be made available to download as an ADS interface. ADS archives are free to use under their Terms and Conditions.
- The ADS interface will present the data in open formats enabling wider re-use, for example Comma Separated Values (.csv)
- The database will also be published as LOD via the ADS triplestore.
- The ADS archive will include file-level and collection-level metadata
- The main ADS archive will present the raw data to download in common and open formats (e.g. CSV or JPG). The LOD can be queried via a SPARQL client or by using the ADS SPARQL query interface.

2.3. Making data interoperable:

ADS collection-level metadata will incorporate a number of LOD vocabularies to facilitate interoperability, these include:

- Heritage data thesauri for subject terms (http://www.heritagedata.org/)
- Getty Thesaurus of Geographic Names for spatial data
- Library of Congress Subject Headings (LCSH)
- The ADS also record spatial data to be compliant with the GEMINI metadata standard
- UB will participate in this task for Catalan and Spanish vocabularies
- UNIPI will contribute with southern-European vocabularies
- UCO with German terminology.

In order to ensure interoperability between resources in different languages, multilingual controlled vocabularies will need to be incorporated into the database. Work in this area for the archaeological domain is being carried out by the EU Infrastructures funded ARIADNE project, which can subsequently be incorporated into this task. As pottery is a subject specialism (depends on the region of production and on the location of the findings), thus sufficient general and language-independent vocabularies do not exist. The project will contribute to create them, and contribute to the larger European resource:

- Heritage data thesauri for subject terms (http://www.heritagedata.org/)
- Getty Thesaurus of Geographic Names for spatial data
- Library of Congress Subject Headings (LCSH)
- The ADS also record spatial data to be compliant with the GEMINI metadata standard
- UB will participate in this task for Catalan and Spanish vocabularies
- UNIPI will contribute with southern-European vocabularies
- UCO with German terminology.

2.4. Increase data re-use (through clarifying licenses):

- The dataset - as delivered via the ADS archive and excluding any material without formal copyright permission - will be freely available to re-use for research purposes as stipulated in the ADS Terms and Conditions of use
- It is anticipated that the data will be available by 2019
- The dataset will be made available by the ADS in perpetuity.
- Details of the ADS Preservation policy and methods of ensuring longevity and security of data can be found in several documents available at: http://archaeologydataservice.ac.uk/advice/preservation

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- ● The costs for depositing the dataset with the ADS, and subsequent resources required to make the dataset publicly available (as a single archive and as LOD) have been included within specific Work Packages within the Archaide project.
- Data management will be overseen by Universitaet zu Koeln and Università di Pisa during the data collection phase, and latterly the ADS as part of the Work Packages to ensure preservation and dissemination.

- The financial costs for ensuring management and presentation of the project dataset by the ADS have been included in the original project design. The impact of the ADS has recently been analysed by an independent study. This project established that the archiving and dissemination of data by the ADS was of significant research and financial value to the wider community.

4. Data security

Data security will be addressed for the period of data collection (1) and deposition of the archive with the ADS (2).

1) The following precautions will be undertaken over the course of the data creation phase:

- This project will follow a rigorous procedures of disaster planning, with (off-site) copies made on a daily, weekly and monthly basis. Backup copies will be validated to ensure that all formatting and important data have been accurately preserved. Each backup will be clearly labelled and its location.

- Periodic checks will be performed on a random sample of digital datasets, whether in active use or stored elsewhere. Appropriate checks will include searching for viruses and routine screening procedures included in most computer operating systems. These periodic checks will be in addition to constant, rigorous virus searching on all files.
- A Preservation Policy: an annual reviewed policy document which alongside detailed descriptions of ADS practice provides an overview of internal procedures for archival policy. This includes an overview of ADS accreditation, migration and backup/off-site storage. The following overview is drawn from this document: "The ADS maintain multiple copies of data in order to facilitate disaster recovery (i.e. to provide resilience). All data are maintained on the main ADS production server in the machine room of the Computing Service at the University of York. The Computing Service further back up this data to tape and maintain off site copies of the tapes. Currently the backup system uses Legato Networker and an Adic Scalar tape library. The system involves daily (over-night), weekly and monthly backups to a fixed number of media so tapes are recycled. All data are synchronised once a week from the local copy in the University of York to a dedicated off site store maintained in the machine room of the UK Data Archive at the University of Essex. This repository takes the form of a standalone server behind the University of Essex firewall. The server is running a RAID 5 disk configuration which allows rapid recovery from disk failure. In the interests of security outside access to this server is via an encrypted SSH tunnel from nominated IP addresses. Data is further backed up to tape by the UKDA.

2) At the end of the project, the dataset will be deposited with the ADS for secure preservation and access into perpetuity. One of the core activities of the ADS is the long term digital archiving of the data that has been entrusted to us. We follow the Open Archival Information System (OAIS) reference model and also have several internal policies and procedures that guide and inform our archiving work in order to ensure that the data in our care is managed in an appropriate and consistent way. These include:

- This project will follow a rigorous procedures of disaster planning, with (off-site) copies made on a daily, weekly and monthly basis. Backup copies will be validated to ensure that all formatting and important data have been accurately preserved. Each backup will be clearly labelled and its location.
- Periodic checks will be performed on a random sample of digital datasets, whether in active use or stored elsewhere. Appropriate checks will include searching for viruses and routine screening procedures included in most computer operating systems. These periodic checks will be in addition to constant, rigorous virus searching on all files.
- A Preservation Policy: an annual reviewed policy document which alongside detailed descriptions of ADS practice provides an overview of internal procedures for archival policy. This includes an overview

of ADS accreditation, migration and backup/off-site storage. The following overview is drawn from this document: "The ADS maintain multiple copies of data in order to facilitate disaster recovery (i.e. to provide resilience). All data are maintained on the main ADS production server in the machine room of the Computing Service at the University of York. The Computing Service further back up this data to tape and maintain off site copies of the tapes. Currently the backup system uses Legato Networker and an Adic Scalar tape library. The system involves daily (over-night), weekly and monthly backups to a fixed number of media so tapes are recycled. All data are synchronised once a week from the local copy in the University of York to a dedicated off site store maintained in the machine room of the UK Data Archive at the University of Essex . This repository takes the form of a standalone server behind the University of Essex firewall. The server is running a RAID 5 disk configuration which allows rapid recovery from disk failure. In the interests of security outside access to this server is via an encrypted SSH tunnel from nominated IP addresses. Data is further backed up to tape by the UKDA.

5. Ethical aspects

All research conducted by University of York staff will be performed in accordance with the Code of practice and principles for good ethical governance.

6. Other

The project Data Management Plan (DMP) presented here is based upon existing internationally agreed procedures and recommendations as outlined in the Archaeology Data Service / Digital Antiquity Guides to Good Practice, as well as specific Digital Preservation based standards including the DCC checklist and handbook of the Digital Preservation Coalition.

In addition to this required format, it was also thought beneficial to have a separate instructive document to guide subsequent Workpackages of the ArchAIDE project and designed to cover practical and technical elements not contained in the online tool. The following recommendations presented here are based upon existing internationally agreed procedures and recommendations as outlined in the Archaeology Data Service / Digital Antiquity *Guides to Good Practice* (Archaeology Data Service/Digital Antiquity 2011), as well as specific Digital Preservation based standards (DCC 2013; Digital Preservation Coalition 2016).

This document covers guidance over the lifetime of the project, from considerations during data collection, deposition with the ADS, and finally preservation and access at the ADS.

6.1. Defining the data to be archived

As defined in Section 3 of this document, the ArchAIDE database is in effect two entities:

- The reference database
- The results database

The reference database will contain a number of digital and digitised catalogues of pottery typologies, and at the end of the project cycle will form a coherent static resource. The results database is intended to form a dynamic user-driven dataset for incorporation based on field and laboratory investigative and reporting workflows. The final ArchAIDE project archive should consist of the reference database and data produced by the application during the project lifetime.

6.2. Data Collection (pre-archiving)

The following Section covers guidelines and recommendations for the period of data creation. It is inherently linked with the formal handover of the archival dataset to the ADS (4.4), and that section should be consulted for specifications on file formats and metadata. During data creation, it is anticipated that the following guidance will be used.

6.2.1. Digitisation

Although a significant amount of data created by the project will be born-digital, a proportion will also be digitised from physical sources. If digitisation is undertaken, a number of organisations and guidelines exist which provide substantial guidance on undertaking digitisation. JISC Digital Media provides a wide range of advice on digitising existing images.

6.2.2. Version Control

Strict version control will be observed. Primarily through the use of

- File naming conventions
- Standard headers listing creation dates and version numbers
- File logs

Versions that are no longer needed will be removed after ensuring that adequate backup files have been created.

6.2.3. File Structures + Naming

Files will be organised into easily understandable directory structures. By following a logical data structure throughout the project, will result in less time preparing data for archiving at the end of the process. Adherence to a predefined file structure will also reduce data loss and it provide files with an absolute location. An example structure is included below; please note that this is not model is used as an example of a clear structure and is not proscriptive.

File naming will be considered from the very outset of a project. Every effort will be made to make file names both descriptive and unique. The following conventions will be used at all times:

- File names should use only alpha-numeric characters (a-z, 0-9), the hyphen (-) and the underscore (_). No other punctuation or special characters should be included within the filename.
- A full stop (.) should only be used as a separator between the file name and the file extension and should not be used elsewhere within the file name.
- Files must have a file extension to help the ADS and future users of the resource determine the file type.
- Lower case characters should be used, and ensure that supplied documentation accurately reflects the case of your filenames.

Some examples would thus be:

- siteid_artefactid_drawing_042.tif
- siteid_artefactid_photograph_012.tif
- siteid_artefactid_model_131.xyz

6.2.4. Secure backup

Backup is the familiar task of ensuring that there is an emergency copy, or snapshot, of data held somewhere other than the primary location. This project will follow a rigorous procedures of disaster planning, with (off-site) copies made on a daily, weekly and monthly basis. These are important in the lifespan of the project, but are not the same as long-term archiving because once the project is completed and its digital archive safely deposited, the action of backing up will become unnecessary. Backup copies will be validated to ensure that all formatting and important data have been accurately preserved. Each backup will be clearly labelled.

6.2.5. Periodic checking for viruses and other issues

Periodic checks will be performed on a random sample of digital datasets, whether in active use or stored elsewhere. Appropriate checks will include searching for viruses and routine screening procedures included in most computer operating systems. These periodic checks will be in addition to constant, rigorous virus searching on all files.

6.3. Archiving with the ADS

At the end of the project, the defined dataset (see 4.2) will be deposited with the ADS for secure preservation and access into perpetuity.

6.3.1. Selection and retention

Through adherence to the guidelines on version control it is hoped that little time should be required for a review of data to be submitted to the ADS. However, a review should be undertaken to ensure that the archive does not contain:

- Duplicates
- Working or backup versions of files
- Correspondence (emails or letters) or informal notes generated over the course of the project (note that if files explain other files within the archive they should be considered as metadata and included)
- Any extraneous or irrelevant materials

6.3.2. File formats

The following formats should be used for deposition of the archive with the ADS. More detail on each datatype is included in the specific sections below

Data type	**File format**	**Notes**
Database	Each table or object should be exported as: • Comma Separated Values (.csv)	UTF-8 encoding should be used if table contain non-ascii characters
Raster images	All raster images should be supplied in any of th following formats: • Uncompressed Baseline TIFF v6 (.tif) • Portable Network Graphic (.png) • Joint Photographic Expert Group (.jpg) • JPEG 2000 (.jp2)	Should be used for photographs and fla drawings. TIF is the ADS preferred forma but others are accepted
Vector images	Scalable Vector Graphics (.svg)	An open standard, XML-based format used t describe 2D vector graphics and develope by the W3C
Computer-Aided Design	AutoCAD (.dwg or .dxf) version 2010 (AC1024)	
3D models	Wavefront OBJ (.obj) X3D (.x3d) Polygon File Format (.ply) Uncompressed Baseline TIFF v6 (.tif)	OBJ, X3D or PLY are acceptable for 3D objects. TIF or DNG should be used for an photographs used for the generation of mode textures

	Digital Negative (.dng)	
Documents	Microsoft Open XML (.docx) OpenDocument Text (.odt)	Either format can be used

6.3.3. Metadata

All files should be accompanied by suitable metadata for that specific metadata type. The ADS has specific guidance and templates for metadata available on its website. Individual links to templates are included in the overview of data types presented below.

6.3.4. Database files

Databases are to be deposited as CSV files – usually as flat exports from the database software being used. For the purposes of the ADS, the core of the database is the data tables along with documentation and metadata describing the contents of and relationships between tables. The order or layout of the columns and rows may also be of significance, but forms, reports, queries and macros are not seen as core data and are therefore often not preserved.

6.3.5. General comments

It is recommended that certain checks be made prior to deposition with the ADS.

- Tables: although it should be assumed that databases should be migrated in their entirety, an assessment should be made in order to establish which tables should be migrated. Tables in the databases used to temporarily store data are not needed for preservation.
- Formulae, Queries, Macros: if the file contains formulae or queries that need to be preserved in their own right then these need to be identified, as migrated versions of the may only preserve the actual values calculated by the functions and not the functions themselves. Queries may need to be preserved separately and documented within a text file so functionality can be recreated at a later date.
- Comments or Notes: as with macros and formulae, the migration process may not save comments or text notes added to a file. Before migration, comments will need to be stored in a separate text file with a clear indication of which file and cell the comment relates to.
- Special Characters: The database may contain special or foreign characters such as ampersands, smart quotes or the em dash ("—") which interfere with the export and subsequent display of the data. Foreign characters which will often not export to a basic text file unless a specific character set (e.g. UTF-8) is specified.
- Links: it is important that the relationships between tables are understood, documented (see below) and are correct (checks can be made to ensure that duplicate or orphan records aren't present). If the database contains links to images, then checks should be made to ensure that these filenames are stored correctly.

6.3.5.1 File metadata

- A template for database metadata can be downloaded from the ADS website here: http://archaeologydataservice.ac.uk/attach/FilelevelMetadata/ADS_database_metadata_template.ods
- An entity relationship model should also be included.

6.3.6. Raster images

The following precautions should be made when creating or converting raster images:

- Image Size and Resolution - conversions should ensure that the original resolution and image size remains the same in the preservation file format. In addition, it is important that, when converting files to a new format, lossy compression is not applied to the image.
- Bit depth and Colour space - converted files should ensure that the bit depth and colour space of the original image are supported in preservation formats and that images are not degraded when converted.

Although these properties are components of all image formats it is important to ensure that these properties remain the same/retain the same values when converting files to archival formats.

In addition, embedded metadata such as EXIF and IPTC can also be seen in certain cases as a significant property of an image and, where relevant, should be preserved with the file or exported to a separate plain or delimited text or XML file to be stored alongside the image. Although it is possible to preserve JPEG EXIF within the TIFF tag structure it is better held in a separate file, avoiding the risk of loss or corruption during later migration and making the metadata more easily accessible. Extraction of EXIF fields is relatively straight forward, with a number of free tools available.

6.3.5.2 File metadata

- A template for raster image metadata can be downloaded from the ADS website here: http://archaeologydataservice.ac.uk/attach/FilelevelMetadata/ADS_raster_metadata_template.ods

6.3.7. Vector images + CAD

Vector images and CAD models should be deposited as either SVG or DWG. Unlike common raster images such as photographs, many vector images are derived from data created or held in other applications such as CAD or GIS (which in turn is often derived from a range of data collection techniques such as geophysical survey or laser scanning). It is advised that if an image is derived from another dataset then preservation of the original file should take precedence over the derived image.

6.3.5.3 File metadata

- A template for raster image metadata can be downloaded from the ADS website here: http://archaeologydataservice.ac.uk/attach/FilelevelMetadata/ADS_vector_metadata_template.ods

6.3.6 Storage at the ADS

All research data collected and generated during the project will be managed securely during the project lifetime, made available as Open Access data by the project end, and securely preserved in the ADS repository into perpetuity. The ADS follows the Open Archival Information System (OAIS) reference model, and have several internal policies and procedures that guide and inform archiving work in order to ensure that the data in our care is managed in an appropriate and consistent way.

All data will be documented in the ADS Collections Management System, an Oracle-based system, held on University of York servers, with a secure off-site backup held in the UK Data Archive at the University of Essex. During the lifetime of the project all partners will maintain current working data on their own secure systems with weekly backup to external hard drives.

References

Archaeology Data Service / Digital Antiquity (2011). *Guides to Good Practice*. Available at http://guides.archaeologydataservice.ac.uk/

DCC (2013). *Checklist for a Data Management Plan. v.4.0*. Edinburgh: Digital Curation Centre. Available at http://www.dcc.ac.uk/resources/data-management-plans

Digital Preservation Coalition (2016). *Digital Preservation Handbook, 2nd Edition*. Available at http://handbook.dpconline.org/