Plan Overview

A Data Management Plan created using DMPonline

Title: ArchAIDE (Horizon 2020 DMP)

Creator: Tim Evans

Principal Investigator: Tim Evans (Archaeology Data Service) tim.evans@york.ac.uk, MARIA LETIZIA GUALANDI

Data Manager: Tim Evans (Archaeology Data Service) tim.evans@york.ac.uk

Affiliation: University of York

Template: Horizon 2020 DMP

Project abstract:

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.

ID: 12379

Last modified: 04-06-2019

Grant number / URL: 693548

ArchAIDE (Horizon 2020 DMP) - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

- 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 database will act as a publicly available reference resource.
- 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 final archive is estimated to consist of a maximum of 100Gb of data.
- 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:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

- 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 resouce 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 aslso 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 Dublic 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:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

- The main output of the project will be the project database. This database will be archived with the Archaeology Data Service (ADS). This database will be made available to download as an the 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:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

ADS collection-level metadata will incorporate a number of LOD vocabualries 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

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:

- 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):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

- The dataset - as delivered via the ADS archive - will be freely available to re-use for research purposes as stipulated in the ADS Terms and Conditions of use
- It is anticapted that the data will be avilable by XXXx
- The dataset will be made available by the ADS in perpituity. Details of the ADS Preservation policy and methods of ensuring longevity and security of data can be found in several documents available on their website

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

- 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 daata by the ADS was of significant research and financial value to the wider community.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Data security will be adressed for the period of data collection (1) and deposition of the archive with the ADS (2).
1) The following precuations will be undertaken over the course of the data creation phase:

- This project will follow a rigourous 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.

2) At the end of the project, the dataset will be deposited with the ADS for sercure 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:

- 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

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

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

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

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

ArchAIDE (Horizon 2020 DMP) - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The purpose of data collection is to build an application which will assist archaeologists in the recognition and classification of pottery sherds. The application will be built via a database, and populated with test datasets from existing resources and catalogues. The database will thus consist of descriptive data, images, the shape models from thse collections.

Explain the relation to the objectives of the project

The database will function as a tool to act as both a reference and recording mechanism for ceramic assemblages

Specify the types and formats of data generated/collected

The database will consist of:

- Descriptive data (i.e. text): will be archived as Comma Separated Vaules (.csv) with UTF-8 encoding
- Raster Images (digital photographs): will be archived as uncompressed TIFF (v6)
- Raster Images (scans of drawings): will be archived as uncompressed TIFF (v6)
- Vector Images: will be archived as either AutoCAD DWG (2010) or Scaleable Vector Graphics (.svg)
- 3D models: will be archived in Polygon File Format (.ply) or nexus format

Any other supporting data or documentation will be archived in the following formats:

- Documents: as either Microsoft Open XML Strict (.docx) or OpenDocument Text (.odt)
- Spreadsheets: as Comma Separated Vaules (.csv) with UTF-8 encoding
- Raster Images (digital photographs): as uncompressed TIFF (v6)

Any formats not in this list will be archived in a format recommended by the digital repository (ADS) according to their current guidleines and in reference to the Guides to Good Practice

Specify if existing data is being re-used (if any)

The database will incorporate the following existing resources:

- Roman Amphorae: a digital resource (doi:10.5284/1028192)
- CERAMALEX - A database unlocking Ptolemaic and Roman pottery finds in Egypt
- Medieval Pottery database of the Università di Pisa

The project could also potentially re-use data from other online resources, including:

- Worcestershire On-line Ceramic Database
- Beazley Archive Pottery Database

Specify the origin of the data

The primary data that will be used to populate the database will be derived from the following sources:

- Roman Amphorae: a digital resource (doi:10.5284/1028192)
- CERAMALEX - A database unlocking Ptolemaic and Roman pottery finds in Egypt held by the Universitaet zu Koeln
- Medieval Pottery database of the Università di Pisa

State the expected size of the data (if known)

The size of the Roman Amphorae database (which will be used to seed the resource) is currently 7Gb, with the additional datasets and potential new data (scans, photographs + 3D models) this may be expected to rise significantly. An estimate of 1 terabyte would represent a maximum expected size.

Outline the data utility: to whom will it be useful

The dataset will present a valuable resource for:

- Academic researchers (staff and students at HE institutions) - primarily those interested in comparisons between pottery types and key characteristics of form and fabric. By presenting an authoratative dataset the resource will also appeal to students, or those wishing to know more about a specific type of ceramic.
- The resource will also present a reference collection that can be used by archaeologists actively engaged in fieldwork or post-excavation, enabling them to recognise and classify ceramics as they are recovered from survey or excavation

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

- 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 resouce 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 aslso 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 Dublic 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

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

The ADS archive will be identifiable via a Digital Object Identifier (DOI), registered with Datacite.

Outline naming conventions used

A Digital Object Identifier (DOI) is an alphanumeric string assigned to uniquely identify an object. It is tied to a metadata description of the object as well as to a digital location, such as a URL, where all the details about the object are accessible.

Outline the approach towards search keyword

ADS Collection-level metadata is based on Dublic 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)

Outline the approach for clear versioning

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.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

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 [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

All of the dataset is to be archived with the Archaeology Data Service (ADS). The data will be made available via the ADS website (http://archaeologydataservice.ac.uk/).

Specify how the data will be made available

The data will be made available via the ADS website - both as downloads and as a queryable interface. ADS archives are free to use under their Terms and Conditions (https://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess.xhtml).

- 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.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

No specialist software should be required to access the data archived with the ADS. The downloadable dataset will be presented in open or common formats (for example .csv or .jpg). The specialist interface (built in html and javascropt) will also allow users to interrogate the data via a web-browser. IMages and 3D models will be presented within the interface - also requiring no additional software or plugins for web browsers.

Specify where the data and associated metadata, documentation and code are deposited

The original datset, file-level metadata and collection-level metadata will be deposited with the ADS.

Specify how access will be provided in case there are any restrictions

There should be no formal restrictions to the datset

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies,

standards or methodologies you will follow to facilitate interoperability.

ADS Collection-level metadata is based on Dublin Core (DC) elements. DC.Subject terms are based on a number of Linked Open Data (LOD) vocabualries to facilitate interoperability, these include:

- Heritage data thesauri for subject terms (http://www.heritagedata.org/). 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)
- 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

In order to ensure interoperability between resources in different languages, multilingual controlled vocabularies will be incorporated into the database. Similar work in the archaeological domain has already been carried out by the EU Infrastructures funded ARIADNE project, mapping country or data centre specific chronologies, object and monument terms to a central neutral spine - the Art and Architecture Thesuarus of the Getty Research Institute.

Following the success of this initiative for ARIADNE, ArchAIDE will use a similar methodology and use the Getty AAT to build a neutral spine of terms specific to ceramic recording. These include:

- Sherd type (for example "rim")
- Form (for example "plate")
- Decoration type (for example "incised")
- Decoration colour (for example "blue")

Project partners will then identify specific terms used within their national or regional catalogues and map them to those neutral concepts.

- UB will participate in this task for Catalan and Spanish vocabularies
- UNIPI will contribute with southern-European vocabularies
- UCO with German terminology.
- University of York for UK terminologies
- An independent ceramic specialist has also contributed an existing thesaurus of English-French terms

The use of the AAT terms will not only allow a linguistic mapping to be incorporated within the reference database and public facing application, but also a conceptual mapping that will allow for differences in terminologies to be overcome. To explain this last point, archaeologists in different countries may have different appreciations of what is a "plate" or "platter". However, in the AAT both terms are hierarchically below a broader term "vessels for serving and consuming food". The database and user interface can use this knowledge organisation to allow the ArchAIDE application to search on very specific terms (such as "plate"), but then to return other results that also map to broader parent terms so as not to omit results based on a subjective and personal appreciation of what an object is called.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

One of the aims of the project is to create a cross-European dataset that accounts for variation in recording of ceramics (for example form or fabric) and to map these to a standard vocabulary (created

by the project). One of the long term aims of the project is that this mapping excercise (and resulting dataset) thus acts as a reference tool for future projects

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data will be deposited with the ADS under a standard ADS licence agreement, a copy of this licence can be seen here. This licence permits use of the data for non-commercial purposes. A detailed overview of this policy can be found in the ADS Terms and Conditions of use

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The datset will be made available upon completion of the project. It is planned that this will occur at the completion of the project in 2019.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

The final datset will be available for download and also made accessble as a queryable interface.

Describe data quality assurance processes

Quality assurance is a high priority for the project. During the**collection phase** all data collected and maintained by partners will be subject to standard best practice, as outlined in the ADS/Digital Antiquity Guides to Good Practice. These practices include basic IT good practice on file naming, strict versioning, secure backups (and maintenance of backups), and virus scanning. In addition, all partners creating data will be responsible for ensuring that that quality of material being produced is sufficient to meet the needs of the project. This will include ensuring that scans and other image captures are of the correct detail and quality to be incorporated within the various modelling applications, and that reference information is correctly entered into the ArcHAIDE database. The ArchAIDE database will be maintained by INERA, with data cleaning, enhancement and validation performed by all project partners

Upon completion of the project the data will be deposited with the ADS, who will ensure that file formats are suitable and that all data is adequately documented to ensure **data preservation**. An overview of the ADS ingest process can be found in ADS Ingest Manual

Specify the length of time for which the data will remain re-usable

The data will be archived and disseminated by the ADS in perpetuity. The ADS is a long-standing and

accredited Digital Repository, with a peer reviewed policy on ensuring long-term preservation (http://archaeologydataservice.ac.uk/advice/preservation).

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

The costs for data management (and by extension making the data FAIR) during the data collection phase have been estimated to be minimal, and covered by the existing scheme of works and funds for the relevant work packages.

The main task to be undertaken to ensure data is FAIR, is the deposition of the final dataset with the Archaeology Data Service, which forms Work Package 10 of the ArchAIDE project. Within WP10, the main body of work for archiving is 10.2 Data archiving, which has been broken down - via a calculation of project months assigned to this task to 28,895 Euros. It should be noted that ADS costs are one-off, and cover the managment and preservation of the dataset in perpetuity.

Clearly identify responsibilities for data management in your 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.

Describe costs and potential value of long term preservation

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 daata by the ADS was of significant research and financial value to the wider community.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Data security will be addressed for the period of **Data Collection** (1) and deposition of the archive with the ADS for **Preservation** (2).

1) During Data Collection all partners will adhere to best practice, as outlined in the ADS/Digital Antiquity Guides to Good Practice. In brief, 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.

2) At the end of the project, the dataset will be deposited with the ADS for sercure preservation and access into perpetuity. Once data has been accessioned into the ADS, it falls under a formal Preservation Policy (see below). The ADS Preservation Policy is written as a peer-reviewed document as part of the organisaitons commitment to being a Trusted Digital Repository, and thus goes into greater detail to cover events pertinent to a long-term archive built upon the Open Archival Information System (OAIS) reference model, and also with 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:

- 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.
- Details of contingencies built as part of Disaster Planning
- Details of contingencies designed to cover transfer of all data to a successor organisation.

It is not anticipated that the final dataset contain any sensitive data. However, deposition of data will be made in reference to the ADS guidlines on the Deposition of Sensitive Digital Data

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Although no ethical issues have been identified, as a matter of course all staff will adhere to the ethical codes and guides to practice of their respective organisations

- University of York (ADS): Code of practice and principles for good ethical governance.
- Tel Aviv Universities ethics policy https://research-authority.tau.ac.il/home/ethics
- University of Barcelona's Code of Good Research Practice http://diposit.ub.edu/dspace/handle/2445/28543
- University of Pisa's ethics code https://www.unipi.it/index.php/statuto-regolamenti/item/1973-codice-etico-della-comunit%C3%A0-accademica

- University of Cologne's Guidelines for Safeguarding Good Academic Practice and Dealing with Academic Misconduct: https://www.portal.uni-koeln.de/sites/uni/PDF/Ordnung_gute_wiss_Praxis_en.pdf
- Elements' ethics code http://elements-arq.weebly.com/ethics-code.html

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

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

ArchAIDE (Horizon 2020 DMP) - Final review DMP

1. Data summary

State the purpose of the data collection/generation

The purpose of data collection has been to build an application which will assist archaeologists in the recognition and classification of pottery sherds. To facilitate this, the project has constructed a MySQL database, and populated with datasets (text and images) from existing digital resources and catalogues, as well as new 3D and vector data generated from both digital and analog sources.

Explain the relation to the objectives of the project

The database functions as a reference resource which can be browsed like a traditional catalogue; the database also acts as the means by which data scanned into the ArchAIDE application is compared.

Specify the types and formats of data generated/collected

The database consists of:

- Descriptive data (i.e. text). Collected in a MySQL instance and archived as Comma Separated Vaules (.csv) with UTF-8 encoding
- Raster Images (digital photographs): collected as JPG and TIF, will be archived as uncompressed TIFF (v6)
- Raster Images (scans of drawings): archived as uncompressed TIFF (v6)
- Vector Images: Scaleable Vector Graphics (.svg)
- 3D models: archived in Polygon File Format (.ply) or nexus format

Any other supporting data or documentation is archived in the following formats:

- Documents: as either Microsoft Open XML Strict (.docx) or OpenDocument Text (.odt)
- Spreadsheets: as Comma Separated Vaules (.csv) with UTF-8 encoding
- Raster Images (digital photographs): as uncompressed TIFF (v6)

Specify if existing data is being re-used (if any)

The database as re-used digital information from the following catalogues:

- Roman Amphorae: a digital resource (doi:10.5284/1028192)
- CERAMALEX - A database unlocking Ptolemaic and Roman pottery finds in Egypt
- Medieval Pottery database of the Università di Pisa

The database has also re-used a large number of digitised paper catalogues, the full list of which is too large to record here.

Specify the origin of the data

The database as re-used digital information from the following catalogues:

- Roman Amphorae: a digital resource (doi:10.5284/1028192)
- CERAMALEX - A database unlocking Ptolemaic and Roman pottery finds in Egypt
- Medieval Pottery database of the Università di Pisa

The database has also re-used a large number of digitised paper catalogues, the full list of which is too large to record here.

State the expected size of the data (if known)

The final size of the deposited archive is approximately 500Gb.

Outline the data utility: to whom will it be useful

The dataset presents a valuable resource for:

- Academic researchers (staff and students at HE institutions) - primarily those interested in comparisons between pottery types and key characteristics of form and fabric. By presenting an authoratative dataset the resource will also appeal to students, or those wishing to know more about a specific type of ceramic.
- The resource also acts as a reference collection that can be used by archaeologists actively engaged in fieldwork or post-excavation, enabling them to recognise and classify ceramics as they are recovered from survey or excavation

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

- The final dataset is being archived by the Archaeology Data Service (ADS) as a single collection https://doi.org/10.5284/1050896. Collection-level metadata (based on qualified Dublin Core) has been created, which allows the resouce to be found within the main ADS website. In the future, this metadata will also be supplied to ARIADNEPLUS.
- In addition, a facet of the dataset (multilingual vocabularies) have been published as Linked Open Data (LOD) via the stores within Allegrograph, and published via Pubby and the ADS' SPARQL interface. The Vocabularies can also be viewed via a direct link at http://data.archaeologydataservice.ac.uk/page/archaide/archaide_vocabularies
- The ADS archive is identifiable via a Digital Object Identifier (DOI), registered with Datacite https://doi.org/10.5284/1050896.
- ADS Collection-level metadata is based on Dublic 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)

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

The ADS archive is identifiable via a Digital Object Identifier (DOI), registered with Datacite: https://doi.org/10.5284/1050896.

Data published as LOD by the ADS have persistent unique identifiers, for example http://data.archaeologydataservice.ac.uk/page/archaide/catalan/type_form/Olla

Outline naming conventions used

A Digital Object Identifier (DOI) is an alphanumeric string assigned to uniquely identify an object. It is tied to a metadata description of the object as well as to a digital location, such as a URL, where all the details about the object are accessible.

Within the archive, file names are based on the following conventions:

- File names only alpha-numeric characters (a-z, 0-9), the hyphen (-) and the underscore (_). No other punctuation or special characters is included within the filename.
- The underscore character is used to to imply a space within the file name.
- A full stop (.) is only be used as a separator between the file name and the file extension and is not used elsewhere within the file name.
- Individual file names, regardless of file structure, are unique within the dataset.
- File names are consistent. Descriptive names are used. A **descriptive file name** helps explain the contents of the file.

Outline the approach towards search keyword

ADS Collection-level metadata is based on Dublic 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 and ARIADNEPLUS projects these terms have also been mapped to the Ariadne Dataset Catalogue Model (ACDM see https://ariadne-infrastructure.eu/)

Outline the approach for clear versioning

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.

Version control the baseline dataset was ensured via a project GIT.

Specify standards for metadata creation (if any). If there are no standards in your

discipline describe what metadata will be created and how

As outlined above, the final archive resides with the ADS with metadata compiled to their standards, based on DC terms. Existing heritage thesauri are used for the recording of subject terms.

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

During the course of the project it became apparant that a large part of the *digitised* corpus used to populate the reference database could not be made openly available due to ongoing copyright; in nearly all cases clearance to submit this data within an open archive was impossible to ascertain.

The final archive made available via the ADS thus represents only the digital material where re-use (under an open licence) has been permitted. This does however inlcude a large amount of derived information, for example the vector (SVG) and 3D models created from the Roman Amphora Database.

Specify how the data will be made available

The data is being made available as an ADS archive https://doi.org/10.5284/1050896 which is made available via the ADS website under a CC BY 4.0 Creative Commons Attribution 4.0 International License.

- The ADS interface will present the data in open formats enabling wider re-use, for example Comma Separated Values (.csv)
- The multilingual vocabularies used in the database have also be published as LOD via the ADS triplestore.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

No specialist software should be required to access the data archived with the ADS. The downloadable dataset will be presented in open or common formats (for example .csv or .jpg) which can be opened and used in a variety of common and open source softwares.

The specialist interface for display of 3D models (built in html and javascript) will also allow users to interrogate the data via a web-browser. Images and 3D models will be presented within the interface - also requiring no additional software or plugins for web browsers.

Specify where the data and associated metadata, documentation and code are deposited

The original datset (subject to copyright clearance), file-level metadata and collection-level metadata has been deposited with the ADS.

Specify how access will be provided in case there are any restrictions

There should be no formal restrictions to the datset archived with the ADS, other than those stipulated under the CC BY 4.0 Creative Commons Attribution 4.0 International License. For the data not being deposited access will be via the ArchAIDE application only.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

ADS Collection-level metadata is based on Dublin Core (DC) elements. DC.Subject terms are based on a number of Linked Open Data (LOD) vocabularies to facilitate interoperability, these include:

- Heritage data thesauri for subject terms (http://www.heritagedata.org/). As part of the ARIADNE and ARIADNEPLUS projects these terms have also been mapped to the ARIADNE Dataset Catalogue Model (ACDM see http://portal.ariadne-infrastructure.eu/about)
- 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

In order to ensure interoperability between resources in different languages, multilingual controlled vocabularies will be incorporated into the database. Similar work in the archaeological domain has already been carried out by the EU Infrastructures funded ARIADNE project, mapping country or data centre specific chronologies, object and monument terms to a central neutral spine - the Art and Architecture Thesuarus of the Getty Research Institute.

Following the success of this initiative for ARIADNE, ArchAIDE have followed a similar methodology and used the Getty AAT to build a neutral spine of terms specific to ceramic recording. These include:

- Sherd type (for example "rim")
- Form (for example "plate")
- Decoration type (for example "incised")
- Decoration colour (for example "blue")
- The characteristics of the rim
- The characteristics of the neck
- The characteristics of the shoulder
- The characteristics of the body
- The characteristics of the base
- The characteristics of the handle

Project partners and associated have identified specific terms used within their national or regional catalogues and mapped them to those neutral concepts. The final mappings were reviewed, to reflect and then incorporate any misunderstandings or inconsistencies. At a later date a contribution to my 2017 ArchAIDE blog by Eleni Schindler Kaudelka drew attention to previous work in this sphere by Caroline Sourzat, and subsequently a copy of her Master's thesis was used to add French terms, and enhance the German terminologies. A further contribution of an e-print of an article from the Journal of Roman Pottery Studies from Nicholas Cooper, an attendee of the December 2017 multiplier event, provided a basis for a mapping in Dutch. This was edited and refined by Leontien Talboom, a Digital

Archivist at the ADS (UoY) and a native Dutch speaker. In late 2018, a mapping in Portuguese has also recently been contributed by Guilherme DAndrea Curra. At the end of this phase, a total of 1338 mappings in seven European languages had been completed.

The mapped vocabularies have been uploaded to the ArchAIDE reference database as a reference resource and are being used for the manual and automated (via text recognition) recording of paper and digital ceramic catalogues. The vocabularies will also be the spine which will allow users of the public application to cross-search the reference database in their own language, and as noted to reconcile differences in understanding or classification

As a separate phase of work to support the reference database and ArchAIDE application and facilitate a wider interoperability, the project team have published the vocabularies and their mappings to AAT as Linked Open Data (http://data.archaeologydataservice.ac.uk/page/archaide/archaide_vocabularies).

In addition to the above, the project database also uses existing vocabularies to denote location, principally Geonames and the Getty Thesaurus of Geographic Names.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

As noted above, collection-level and file-level metadata will use ontologies common to Heritage practitioners (http://www.heritagedata.org/), as well as international standards such as LCSH and Getty. Use of international vocabualries such as Getty and LCSH will allow at least a basic interoperability with non-heritage disciplines.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data has been deposited with the ADS, and licenced under a CC BY 4.0Creative Commons Attribution 4.0 International License.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The mulitlingual vocabualries have already been made archived and made publicly available; the database and associated data is currently being ingested into the archive, and will be made accessible over Summer 2019.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

The final datset is/will be available for download under a CC BY 4.0 Creative Commons Attribution 4.0

International License

Describe data quality assurance processes

During the **collection phase** all data collected and maintained by partners has been guided by the ADS/Digital Antiquity Guides to Good Practice. These practices include basic IT good practice on file naming, strict versioning, secure backups (and maintenance of backups), and virus scanning. In addition, all partners creating data have been responsible for ensuring that that quality of material being produced is sufficient to meet the needs of the project (with a particualr emphasis on image production).

The ArchAIDE database has been maintained by INERA, with data cleaning, enhancement and validation performed by all project partners to esnure accuracy of records.

Upon completion of the project the data has been deposited with the ADS, who will ensure that file formats are suitable and that all data is adequately documented to ensure **data preservation**. An overview of the ADS ingest process can be found in ADS Ingest Manual

Specify the length of time for which the data will remain re-usable

The data will be archived and disseminated by the ADS in perpetuity. The ADS is a long-standing and accredited Digital Repository, with a peer reviewed policy on ensuring long-term preservation (http://archaeologydataservice.ac.uk/advice/preservation).

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

The costs for data management (and by extension making the data FAIR) during the data collection phase have been estimated to be minimal, and covered by the existing scheme of works and funds for the relevant work packages.

The main task to be undertaken to ensure data is FAIR, is the deposition of the final dataset with the Archaeology Data Service, which forms Work Package 10 of the ArchAIDE project. Within WP10, the main body of work for archiving is 10.2 Data archiving, which has been broken down - via a calculation of project months assigned to this task to 28,895 Euros. It should be noted that ADS costs are one-off, and cover the managment and preservation of the dataset in perpetuity.

Clearly identify responsibilities for data management in your 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.

Describe costs and potential value of long term preservation

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 daata by the ADS was of significant research and financial value to the wider community.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Data security has been addressed for the period of **Data Collection** (1) and deposition of the archive with the ADS for **Preservation** (2).

1) During Data Collection all partners have adhered to best practice, as outlined in the ADS/Digital Antiquity Guides to Good Practice. In brief, the following precautions have been undertaken over the course of the data creation phase:

- This project has followed a rigorous procedures of disaster planning, with (off-site) copies made on a daily, weekly and monthly basis. Backup copies have been validated to ensure that all formatting and important data have been accurately preserved. Each backup has been clearly labelled and its location.
- Periodic checks have been performed on a random sample of digital datasets, whether in active use or stored elsewhere. Checks have included searching for viruses and routine screening procedures included in most computer operating systems.

2) At the end of the project, the dataset has been deposited with the ADS for sercure preservation and access into perpetuity. Once data has been accessioned into the ADS, it falls under a formal Preservation Policy (see below). The ADS Preservation Policy is written as a peer-reviewed document as part of the organisaitons commitment to being a Trusted Digital Repository, and thus goes into greater detail to cover events pertinent to a long-term archive built upon the Open Archival Information System (OAIS) reference model, and also with 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:

- 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.
- Details of contingencies built as part of Disaster Planning
- Details of contingencies designed to cover transfer of all data to a successor organisation.

The final dataset does not contain any sensitive data. However, ADS ingest includes guidelines on the Deposition of Sensitive Digital Data

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Although no ethical issues have been identified over the course of the project, as a matter of course all staff have been notified of, and have adhered to the ethical codes and guides to practice of their respective organisations

- University of York (ADS): Code of practice and principles for good ethical governance.
- Tel Aviv Universities ethics policy https://research-authority.tau.ac.il/home/ethics
- University of Barcelona's Code of Good Research Practice http://diposit.ub.edu/dspace/handle/2445/28543
- University of Pisa's ethics code https://www.unipi.it/index.php/statuto-regolamenti/item/1973-codice-etico-della-comunit%C3%A0-accademica
- University of Cologne's Guidelines for Safeguarding Good Academic Practice and Dealing with Academic Misconduct: https://www.portal.uni-koeln.de/sites/uni/PDF/Ordnung_gute_wiss_Praxis_en.pdf
- Elements' ethics code http://elements-arq.weebly.com/ethics-code.html

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

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