Plan Overview

A Data Management Plan created using DMPonline

Title: ArchAIDE (Horizon 2020 DMP)

Creator: Tim Evans

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 alongsidedetailed 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 theCode 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