Data Management Plan

Version 1.0

Grant Agreement number	809994
Programme call	ERC Synergy Grant 2018
Funding Agency	European Union's Research and Innovation Programme Horizon 2020
Action Acronym	DHARMA
Action Title	The Domestication of "Hindu" Asceticism and the Religious Making of South and Southeast Asia
Corresponding Principal Investigator	Emmanuel FRANCIS (CEIAS, UMR 8564, EHESS & CNRS)
Other Principal Investigators	Arlo GRIFFITHS (EFEO), Annette SCHMIEDCHEN (UBER)
Project start date	May 1st, 2019
Duration	72 months
Contacts	arlo.griffiths@efeo.net adeline.levivier@efeo.net axelle.janiak@ehess.fr

Versioning History

Author(s)	Version	Changes	Date
Axelle Janiak & Adeline Levivier	0.1	Creation of the file and redaction of the first draft	2019-10
Emmanuel Francis, Arlo Griffiths, Emmanuelle Morlock, Vincent Paillusson & Annette Schmiedchen	0.2	Review of the draft	2019-10
Emmanuel Francis, Arlo Griffiths, Axelle Janiak, Adeline Levivier & Annette Schmiedchen	1.0	Final version	2019-10

Table of Content

List of Abbreviations

API	Application Program Interface
ARK	Archival Resource Key
CC-BY	Creative Commons Attribution 4.0 licence
CC0	Creative Commons Zero 1.0 licence
CEIAS	Centre d'Études de l'Inde et de l'Asie du Sud
CINES	Centre Informatique National de l'Enseignement Supérieur
CNRS	Centre National de la Recherche Scientifique
DCAT-AP	Application profile for data portals in Europe
DMP	Data Management Plan
DOI	Digital Object Identifier
EFEO	École française d'Extrême-Orient
EHESS	École des Hautes Études en Sciences Sociales
FAIR	Findable, Accessible, Interoperable and Reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
HTTP	HyperText Transfer Protocol
IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
IPTC	International Press Telecommunications Council
JSON	JavaScript Object Notation
OAI-PMH	Open Archive Initiative Protocol for Metadata Harvesting
ODD	One Document Does it all
PTM	Polynomial Texture Map
RDF	Resource Description Framework
REST	Representational State Transfer
RTI	Reflectance Transformation Imaging
UBER	Humboldt-Universität zu Berlin
UNO	Università degli Studi di Napoli L'Orientale
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
XSLT	eXtensible Stylesheet Language Transformations

1. Introduction

1.1 Overview of the Project

The religion known today as “Hinduism” is a major world religion and the main religion of the world’s largest democracy, India. But the history of “Hindu” institutions is not limited to India. The DHARMA project will study the history of “Hinduism” in comparative perspective, focusing on the period from the 6th to the 13th century. During this period, the Bay of Bengal served as a maritime highway for intense cultural exchange. The resulting process of “Indianisation”, marked notably by the spread of “Hinduism”, of an Indian writing system and of India’s sacred language Sanskrit, impacted large parts of South and Southeast Asia.

The Sanskrit word DHARMA can designate the cosmic law that is upheld both by gods and humans. But it is also often used to refer to any of the numerous temple-related foundations made to support this law. The DHARMA project seeks to understand the process of “institutionalisation” of “Hinduism” by investigating the roles of various agents, from kings and noblemen to priests, monks and local communities. It emphasizes social and material contexts of “Hinduism”, which requires a multi-regional, multi-scalar and multi-disciplinary methodology, to forge a real synergy of scholarship on premodern South and Southeast Asia.

The approach will be based on the correlation and contextualisation of written evidence from inscriptions and manuscripts and material evidence from temples and other kinds of archaeological sites. The project will be carried out in four task-forces. Three regional task forces will focus, respectively, on the inscriptions and archaeological sites of the Tamil-speaking South of India (A), of Central through North-eastern South Asia into what is today Myanmar (B), and of mainland plus insular Southeast Asia (C). A fourth, transversal task-force (D), will focus on textual material transmitted in manuscript form.

Inscriptions are the main sources for the history of premodern South and Southeast Asia. But they are not all accessible, even less so in a machine-actionable format. For the large-scale comparative research undertaken, making as much as possible of South and Southeast Asian epigraphy available, in a digital database,[1] is therefore a core objective of the project. South and Southeast Asian manuscripts, normally written on palm-leaf, preserve a rich textual archive relevant to the history of “Hinduism”. The project will produce editions with translations of texts that have so far remained unpublished. These include descriptions of religious practices, as well as prescriptions that deal both with lay religiosity and with religious life in temples and monasteries. As for archaeological evidence, surveys and excavations will be conducted at sites which are known to be rich in data. These will make it possible to confront the findings in the inscriptions and texts with the material record.

1.2 Data Management Plan

The DHARMA project is participating in the European Commission’s Open Research Data Pilot. This implies giving access to data free of charge but also to enable reuse of research data that have been used to validate the results presented in scholarly publications as well as the publications themselves.

[1] The DHARMA-base has to be understood as a generic designation of the project’s digital ecosystem. It will be defined more precisely at a later stage of the project.

This document establishes the management plan for data collected and created for the project in order to facilitate the reuse of the data and their accessibility in the long term. The purpose of this document is to plan the dissemination of those data following best practices to make the data findable, accessible, interoperable and reusable (FAIR) and to prevent any future loss of data. Indeed, the DMP specifies how the data will be handled during and after the project and takes into consideration such aspects as the collection, storage, security and retrieval of data, as well as the costs involved in these various stages of the data lifecycle. The present document follows the version 3.0 of the *Guidelines on FAIR Data Management in Horizon 2020*,[2] as well as the recommendations for research data management established by the Direction de l'information scientifique et technique (DIST) of the CNRS.[3] It is also compliant with UBER's research data management policy established in 2014.[4] The EFEO doesn't have a data management policy yet (though one is in preparation).

This DMP for the DHARMA project will evolve during the lifespan of the project and will be adapted as and when the need arises due to the development of the project. It is currently in its first version and will be submitted to the ERC as formal deliverable of the project due by the end of October 2019. Versions updated during the course of the project will be resubmitted to the ERC in the case of significant changes, as part of the Periodic Reports.

2. Summary of Data

This section provides a description of the data that are at this stage of planning expected to be collected and generated during the project. This description includes relevant information about type/format and volume. Transformation processes are described only when an assessment has already been conducted considering use for the project but also long-term preservation.

2.1 Digital Scholarly Editions

We will create and collect editions of epigraphic texts (inscriptions) as well as of texts transmitted in manuscript(s) in the form of XML files encoded following the TEI standard,[5] most often in its EpiDoc specification.[6] We will be mobilizing encoded texts from previous projects undertaken by DHARMA team members, but also files created as part of the project, whether by team members or by an external party, for instance files whose creation we commission with DHARMA funds, as in the case of digitization of printed editions (by OCR or double keying) encoded by subcontractors with a very light set of TEI elements.

All the procedures of validation, conversion and harmonization will be made under the supervision of the project's XML-TEI Data Manager.

---

[2] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] https://ordar.otelo.univ-lorraine.fr/files/ORDAR-1/GuideBP_Otelo_Inist_20170620.pdf

[4] https://www.cms.hu-berlin.de/de/dl/dataman/hu-rdm-guidelines/view

[5] The Text Encoding Initiative (https://tei-c.org/) is a consortium which develops and maintains a standard for encoding texts in digital form using XML.

[6] EpiDoc is a subset of the TEI standard to encode editions of ancient documents. See Tom Elliott, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, et al., 2007-2017, *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*, available at http://www.stoa.org/epidoc/gl/latest/.

Formats	
Collected formats	.txt, .rtf, .doc, .docx and .xml
Working formats	.xml compliant with the latest EpiDoc documentation and TEI P5
Dissemination format	.xml and .html
Archiving format	Each dataset in .xml will be delivered with the DHARMA project's schema in the form of a TEI ODD file[7] (containing the specification of the schema and its documentation)
Storage	
Current storage platforms	GitHub repositories[8]
Dissemination platforms	GitHub repositories, DHARMA-base and Zenodo[9]
Long-term repository	Zenodo via GitHub
Volume	250-500 GB
Process	
Editing	All the files will be encoded in accordance with the project's Encoding Guide.[10] Editing will be done using a text editor. The most common one in the field is the proprietary editor Oxygen XML,[11] but any other open software is possible.
Metadata	The metadata of the text and the text-bearing artefact will be displayed in each <teiHeader>.
Transformation process	For collected digital editions, the encoding will be transformed and validated against the DHARMA schema that we will develop. The conversion will be handled automatically (via XSLT transformations) as much as possible. The features from the DHARMA schema that were not convertible will be encoded manually by the researchers. The file will be checked against the RelaxNG schema and proofread by the researchers.
Conversion tools	XSLT transformations run with Saxon-HE processor. Some of the needed stylesheets can be reused from the freely available TEI and EpiDoc collections.

[7] ODD (https://wiki.tei-c.org/index.php/ODD) is a TEI XML-conformant specification format allowing customization of TEI through literate programming.
[8] GitHub (https://github.com/) is a commercial web hosting service for development of software.
[9] Zenodo (https://zenodo.org/) is an open-access repository developed under the European OpenAIRE program and operated by CERN.
[10] A draft of these guidelines is in advanced state of preparation at the time of the redaction of the first release of this Data Management Plan, and the first release of the Encoding Guide is expected in a matter of weeks.
[11] Oxygen (https://www.oxygenxml.com/) is a complete XML editing software developed by Syncro Soft since 1998.

Validation process	All files created and collected will be validated against a RelaxNG schema. The automatic validation will always be completed by project-internal review of the scholarly contents.
Validation tools	The ODD will be established with Roma.[12]
Rights	
Metadata rights	CC0
Resource rights	CC-BY

2.2 Visual Data

We will collect and create visual reproductions of original text-bearing artefacts in the form of photographic coverage of inscriptions, of estampages of inscriptions, and of manuscripts. To a large extent, we will be collecting visual data that we have already assembled in the course of our previous research projects; in some cases the data have been created by ourselves, in others they have been obtained from third parties (e.g. photos ordered from museums).

Team members will deliver image files to the project's Visual Data Manager along with a transfer form, which will gather all the information needed to identify the content of the files.

The creation of the files, shooting and editing of photos, RTI[13] and photogrammetry, will to a large extent be done by the Visual Data Manager and will normally be in RAW, or at least in TIFF format. If image files are created by another person, whether project researcher, contractor or obtained from an institution such as a museum, they will be considered as collected images and will be delivered to the Visual Data Manager with a transfer form.

Formats	
Collected formats	.jpg, .raw, .tiff, .png, .ptm, .rti, .obj and .ply The settings of the .tiff and .jpg files will follow the standard NISO Z39.87-2002.[14]
Working formats	.jpg, .ptm, .ply, .rti and .obj
Dissemination format	.jpg, .png, .rti and .obj
Archiving format	.tiff, .ptm and .obj (JPEG2000 is also acceptable if that is how any image files will be delivered)
Storage	

[12] Roma (https://roma2.tei-c.org/) is an open-source application developed by the TEI community.
[13] Reflectance Transformation Imaging (http://culturalheritageimaging.org/Technologies/RTI/) is a computational photographic method, developed by Cultural Heritage Imaging. This method captures an object's surface shape and color and enables the interactive re-lighting of the subject from any direction.
[14] https://www.niso.org/publications/ansiniso-z3987-2006-r2017-data-dictionary-technical-metadata-digital-still-images

Current storage platforms	ShareDocs[15] and Didómena[16] for documentation format and Huma-Num Box[17] for heavier formats, such as .tiff and .raw
Dissemination platforms	Displayed on the DHARMA-base and stored in a triplestore database, such as Nakala[18] for each image selected into the database, Zenodo for the published dataset and Didómena for broader sets of images created in the field but not selected into the database.[19] Each repository provides IIIF standard protocol for the interoperability of images.
Long-term repository	CINES[20] through Huma-Num or Zenodo
Volume	1-10 TB range
Process	
Editing	To edit the photos, we will use the software Adobe Lightroom Classic.[21] To generate RTI files, we will use the open-source applications RTIBuilder[22] and RTIViewer.[23] To generate 3D files, we will use the software Agisoft Metashape.[24]
Metadata	A minimal set of metadata will be added to the images in order to be able to identify them. First the members of the project who create image files will have to edit metadata adding the minimal set. The open-source applications to edit metadata that we will use are XnView and ExifTool.[25] Then team members will have to name their images according to the file naming conventions of

[15] Sharedocs is a repository for current documentation to store, share and manage all types of files, made available by the French digital research infrastructure Huma-Num (https://www.huma-num.fr/services-et-outils/stocker).
[16] Didómena (https://didomena.ehess.fr/?locale=en) is the research data repository of the EHESS, one of the Host Institutions of the project.
[17] Huma-Num Box (https://documentation.huma-num.fr/content/23/211/fr/huma_num-box-presentation-globale.html?) is a set of online storage services for heavy datasets offered by Huma-Num.
[18] Nakala (https://www.nakala.fr/) is a repository offered by Huma-Num which provides permanent identifiers, for data and metadata, permanent data access and exposition of metadata through a Triple Store and OAI-PMH . It also provides IIIF services for images.
[19] While Didómena is a service provided by EHESS, an affiliation to EHESS is not mandatory to use it.
[20] The CINES (National Computing Center for Higher Education) is a French public institution, supervised by the French Ministry for Higher Education and Research, in charge of the long-term digital archiving. Note that the CINES enables long-term archiving through the platform PAC (https://www.cines.fr/archivage/nos-solutions-darchivage/pac/workflow-d-archive/), but since 2015, the ministries of Foreign Affairs, Defense and Culture, responsible for State archives and records keeping, are working to create open source software to ingest, manage, preserve and provide long-term access to digital records and archives (http://www.programmevitam.fr/). It is not yet accessible through Huma-Num at the time we submit the first release of this DMP, but we will use it as soon as this becomes possible. If it is not set up before the end of the project, the long-term storage will have to be entrusted to another service provider.
[21] https://www.adobe.com/fr/products/photoshop-lightroom.html
[22] http://culturalheritageimaging.org/What_We_Offer/Downloads/Process/index.html
[23] http://culturalheritageimaging.org/What_We_Offer/Downloads/View/index.html
[24] https://www.agisoft.com/features/professional-edition/
[25] https://www.xnview.com/en/ and https://www.sno.phy.queensu.ca/~phil/exiftool/

	the project[26] and, finally, they will send their files with a form containing all the information necessary for their management.
Metadata extraction	To extract metadata, we will use the open-source software ExifTool, in order to preserve the EXIF/IPTC IIM/XMP[27] metadata in a CSV file.
Transformation process	In a first step, all the files that are not in TIFF format will be converted to this format, with the valid settings expected for archival. In a second step, if necessary, a copy in JPEG[28] format will be created for current documentation.
Conversion tools	To convert JPEG, RAW, and PNG formats, we will use the open-source software XnView, which makes it possible to simultaneously convert, add metadata and change filenames. IPTC metadata will be added to each file according to the guidelines of the project for delivery of image files.[29] The RAW files will be converted, using Adobe Lightroom Classic CC, into JPEG for current records and into TIFF for archiving records. The RAW files will be preserved in order to create new derivatives.
Validation process	The files already in TIFF and JPEG formats will be checked to validate the well-formedness of their format. The files resulting from conversion will systematically be conformant to the chosen formats.
Validation tools	To validate native JPEG and TIFF files, we will use the open-source software JHOVE.[30] If the files are not valid, they will be appropriately converted.
Rights	
Metadata rights	CC0
Resource rights	CC-BY

[26] A draft of these guidelines is in advanced state of preparation at the time of the redaction of the first release of this Data Management Plan, and a first release is expected in a matter of weeks.
[27] Exchangeable Image File Format (https://owl.phy.queensu.ca/~phil/exiftool/TagNames/EXIF.html) is a metadata encoding format to describe images. Information Interchange Model (https://iptc.org/standards/iim/) for International Press Telecommunications Council is a set of metadata used to describe media. Extensible Metadata Platform (https://www.adobe.com/devnet/xmp.html) is a standard ISO data model to describe media, particularly images.
[28] The conversion will be in JPEG: 240 dpi, 8 bits per sample.
[29] A draft of these guidelines is in advanced state of preparation at the time of the redaction of the first release of this Data Management Plan, and a first release of the Guide is expected in a matter of weeks.
[30] Jhove (http://jhove.openpreservation.org/) is an open-source software for file format identification, validation and characterisation tool.

2.3 Bibliographic Data

The bibliographic data will be managed in a Zotero group and often be aggregated from previous projects or else from online catalogs. Single items and bulk import can be handled by researchers themselves, as long as their format is directly sustained by the Zotero system.[31] For other formats, transformation will be necessary and will, as far as possible, be handled by the project's XML-TEI Data Manager.

Formats	
Collected formats	.xml, .docx, .rdf and .xlsx
Working formats	.rdf
Dissemination format	.html and potentially a .xml version
Archiving format	.rdf
Storage	
Current storage platforms	Zotero services
Dissemination platforms	DHARMA-base and Zotero website[32]
Long-term repository	Export in a bibliographic format
Volume	10 GB
Process	
Editing	The data will be collected from online library catalogs as much as possible using the browser plugin Zotero Connector and will be cleaned manually according to the conventions laid down in the project's Zotero Guide.[33]
Metadata extraction	Metadata stored in Zotero can be extracted in bulk with the API in JSON, which means we are able to maintain a certain independence vis-à-vis this specific tool. So the information can be exported to be uploaded in another system if and when necessary, and into the DHARMA-base itself.
Validation process	Review by team members and final validation by PIs.
Validation tools	None for the moment
Rights	
Metadata rights	CC0

[31] Zotero (https://www.zotero.org/) is an open-source software to manage bibliographical data, developed since 2006 by the Center for History and New Media of the George Mason University.
[32] The Group Library is open to the public, but modifications require writing rights.
[33] The first version has been disseminated inside the project. The second version is expected in a matter of weeks and will be published more broadly.

Resource rights	CC-BY

2.4 Archaeological Data

The archeological data will be handled differently from the preceding scenarios. Indeed, each archeological excavation is expected to produce a very large amount of photos (between 1 and 10 TB), all in JPG format, as well as geographical data.[34] All the data created will be handled in the framework of the individual excavation projects since some of them exist for at least 20 years and have their own coherent system. A subset of photos for each excavation campaign will be delivered in RAW/TIFF/PNG format to DHARMA's Visual Data Manager. These photos will be chosen among the ones included in the report and will be the most representative of a given campaign. This selection will be treated according to the procedures described for Visual Data above (§2.2) and each report as Scientific Publication (§2.5).

Inventories of photos for each excavated site will also be delivered, documenting the research material available. The description will specify the provenance, the authors, the excavation campaign, and the license, completed, when appropriate, with elements regarding the authorization negotiated with the concerned authorities, the person or institution in charge of the curation of the collection and the conditions under which the photos may be accessed.

The archaeological subprojects require several professional software applications. Most of them are proprietary and they do not always have satisfactory counterparts in open-source software. For efficiency reasons, use of such proprietary software will be maintained. However, an export in open-source format will be delivered for long-term preservation purposes. Only the final version will be delivered by the archaeologists to DHARMA's Visual Data Manager.

Formats	
Collected formats	.jpg, .raw, .tiff, .png, .ptm, .rti, .obj, .ply and .ai
Working formats	.pdf, .raw, .jpg, .tiff and .svg
Dissemination formats	.pdf, .jpg and .svg
Archiving format	.pdf, .tiff and .svg
Storage	
Current storage platforms	Siem Reap NAS server, PSIR/MOM in Lyon NAS server,[35] own storage of the excavation project in question, and ShareDocs. Any sensitive data will be stored on a closed repository, such as Huma-Num Box.

[34] For now, we don't have an overview about the geographical data which will be collected in the field. The project budget provides for recruitment of a Geographical Data Manager during the second half of the project (2022–2025).

[35] These servers will be handled by the centers and not by DHARMA team. Note that the archaeological team working in Indonesia has not reached a satisfying storage solution yet.

Dissemination platforms	HAL-SHS[36] and ARIADNEplus portal[37]
Long-term repository	CINES through Huma-Num or Zenodo, except for the sensitive data which can not be disseminated or archived.
Volume	1-10 TB range
Rights	
Metadata rights	CC0
Resource rights	CC-BY, but as some data are sensitive, their access will be subject to agreements with the local authorities

2.5 Scientific Publication, Training Material and Technical Documentation

In accordance with Article 29.2 of the *DHARMA Grant Agreement*, an electronic copy of the published version or the accepted peer-reviewed manuscript of all peer-reviewed publications resulting from the project, will be deposited in a repository for scientific publications. Other written work produced by the project, such as training material, presentations, posters, as well as technical documentation, will likewise be deposited in a suitable repository. All relevant repositories will bear the label DHARMA. Deposit will be carried out either by the researchers who have access to them or by the project's European Project Manager, and the metadata will be validated by Visual Data Manager and XML-TEI Data Manager.[38]

Formats	
Collected formats	.docx, .pdf, .odt, .md, .pptx and .tex
Working formats	.docx, .pdf, .odt, .md, .pptx and .tex
Dissemination formats	.pdf and .md[39]
Archiving format	.pdf
Storage	
Current storage	ShareDocs and GitHub

[36] https://halshs.archives-ouvertes.fr/

[37] ARIADNEplus is funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914. It is the extension of the previous ARIADNE Integrating Activity program. Its goal is to index in its registry about 2.000.000 archeological datasets. An agreement will have to be negotiated by the DHARMA project to be able to benefit from dissemination through ARIADNEplus.

[38] Note that if the researcher publishes himself on a repository for scientific publications, he will share the access rights to the co-writers and to the European Project Manager, in order to validate the description of the publication. It is also his responsibility to follow the metadata requirements established by the Visual Data Manager and XML-TEI Data Manager.

[39] Note that some technical documentation will stay in .md depending on the content. Markdown uses a plain-text formatting syntax.

platforms	
Dissemination platforms	HAL-SHS, Zenodo, GitHub and *Hypotheses* on Open-Edition[40]
Long-term repository	CINES through HAL-SHS or Zenodo
Volume	250-500 GB
Process	
Editing	The choice of the tools is left to the author. Note that if modifications are made a version of each final release will be delivered.
Metadata	A metadata file will be delivered to document the files in the repositories.
Transformation process	When necessary, the members of the project will themselves handle the transformation of their publication into PDF before delivering it to the designated person.
Conversion tools	We will use the export feature embedded in the tools used by the researcher so as to simplify our workflow.
Validation process	All the export for preservation will be done in .pdf, and all .pdf files will be checked to validate their well-formedness.
Validation tools	All .pdf files must be in PDF-A format. We will use the open-source software JHOVE to validate its compliance to the format. Non-conform .pdf will be converted.
Rights	
Metadata rights	CC0
Resource rights	CC-BY: if there is a temporary embargo imposed by the publisher, the access to the publication will be restricted during the embargo period, while the metadata of the publication will still be accessible.

3. Adherence to FAIR Principles

3.1 Making the Data Findable

3.1.1 Repositories

Several repositories will be used for short-term and long-term storage of the research data. Some of our repositories are to be understood as workspaces and not as open repositories.

[40] https://dharma.hypotheses.org/

ShareDocs allows us to store administrative documentation as well as current image files while we use Huma-Num Box for heavier files, particularly images. ShareDocs makes it possible to exchange documents and work on them collaboratively in a secure environment.

The GitHub and Zotero services are somewhere in between workspaces and repositories for dissemination, since it is possible to give access to the content, or make it entirely closed to outsiders of the project. We have decided to open the content so it can be read by anyone, but without allowing non-team-members to modify our data.[41]

The technical documentation, transformation stylesheets (XSLT), schemas (Relax NG and Schematron) and codes will be accessible on public repositories hosted by GitHub. All such data are gathered in the GitHub organization labeled "DHARMA". If and when it will be judged interesting for a larger community, some of these routines will be submitted for inclusion in the markup collections on GitHub for Epidoc. Moreover, most of them will be archived in the research data repository Zenodo thanks to the automatic synchronization established between the two services based on webhooks.

Zenodo is developed and hosted on top of existing infrastructure and services at CERN. A Zenodo community called "DHARMA" will be created to gather all the datasets related to the project.[42] Each dataset has its own record and metadata as well as a persistent identifier (DOI) to provide easy upload and citation.[43] In addition, some photographic material will be uploaded on Didómena in a "DHARMA" collection.[44] At corpus level, each dataset uploaded will be given a DOI and documented with the necessary metadata.[45]

All of the project's scholarly publications will be deposited on HAL-SHS in a collection called "DHARMA" that will be created for dissemination purposes. This repository, developed by the CCSC of the CNRS, offers a service for the preservation of research results. It allows attributing index terms and offers metadata templates in which we can record DOIs.[46]

All datasets uploaded into Zenodo,[47] HAL-SHS and Didómena will be preserved for the lifespan of those repositories. In case of any closure, best efforts will be made to integrate the content into any alternative institutional or disciplinary repository.

3.1.2 Provisions for Metadata

All datasets will be described with rich metadata based on the methodologies and interests of the team members. They will be described in order to be compliant with DataCite's Metadata Schema minimum and recommended terms with a few additions to meet requirements set by

[41] Opening the content would have made the validation and control on the data a necessity. We don't have the means to handle such a scenario within the duration of the project.
[42] The total file size limit per record on Zenodo is 50 GB. For our bigger datasets, it might be necessary to request a higher quota.
[43] Note that Zenodo supports DOI versioning which makes it possible to update our data after initial deposit. With this feature, one may cite different versions or all the versions of a record through a specific DOI.
[44] https://didomena.ehess.fr/collections/b5644r887
[45] Note that Didómena does not require delivery of metadata for the datasets deposited. Nevertheless, in our use of this repository, we will ensure delivery of metadata in order to comply with the FAIR principles and be coherent with all the descriptions of our data.
[46] Note that DOIs are not always provided by publishers. Since Zenodo offers DOIs for free, the publications that don't have DOIs will first be deposited in Zenodo in order to obtain a DOI before being deposited in HAL-SHS.
[47] Zenodo is hosted on the servers of the laboratory CERN, which has an experimental programme planned to remain active for the next 20 years at least.

other services, e.g. the Open Access initiative OpenAIRE. The metadata for each digital entity will include information regarding the funding agency, grant ID number and project acronym.

The DOI should always clearly and explicitly be included in the data. This requirement is especially true for datasets which will be deposited in Zenodo and Didómena, since the engines of these two repositories index the metadata and make them searchable right after deposit. The metadata will be sent to DataCite servers for the DOI registration procedure in order to be indexed there as well. They will furthermore be harvestable via the OAI-PMH[48] protocol in its version 2.0.[49] Their initial storage format will be in JSON format, but they can be exported as single item in several bibliographic formats.

The metadata catalog of the inscriptions themselves will be available and searchable in the DHARMA-base. Each inscription and artefact will have a description using a controlled vocabulary system. Several possibilities are currently under study like the standard DCAT-AP vocabulary which is the profile used for the European data portals, based on the recommendations of the W3C.[50]

Each file has its own metadata. For text editions, they will be contained and expressed in the relevant XML file, while for photographs, they will be embedded in the image files and listed in CSV files in each folder.

3.1.3 File Naming Conventions and Versioning

Our policy is to create consistent and predictable but still user-friendly dataset names, that must also be valid as URL.[51] We take care to provide names that will be understandable even for outsiders of the project, every file name being ready for dissemination and archiving from the start thanks to the following pattern:

DHARMA_<subject>_<datatype>_<version>_<date>.<extension>

Note that the <subject> element can be structured in several different ways depending on the file's content, while <version> and <date> are not mandatory elements.

Important datasets or documentation evolve and grow. To address the need to control the process of modification of those files while also maintaining a stable base version in the case of data that have already been published, the main releases will be kept in a static version and deposited in the appropriate repository. In parallel a dynamic version will still be hosted in the relevant project repository. Ideally, both versions would be findable. In reality, this will be achieved only for the resources in plain text (.txt) and XML markup (.xml), since they are stored on open-source platforms.

For the files handled with the automated versioning system Git, we will have a full record of the revision history. The files managed with other tools will be documented with a revision history table.[52] The main releases and versions of those files will be recorded and made available thanks to the versioning mechanism available in the data repositories.

---

[48] The Open Archives Initiative Protocol for Metadata Harvesting (https://www.openarchives.org/pmh/) is a low-barrier mechanism for repository interoperability.

[49] Zenodo is registered as an OAI Data provider (https://zenodo.org/oai2d). The service is accessible at the base URL. It is also registered with the Directory of Open Access Repositories (OpenDOAR) and the Registry of Research Data Repositories (re3Data).

[50] https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

[51] They are described in a guide on DHARMA File Naming Conventions that is nearly finished at the time we submit the first version of this DMP and will be made accessible on GitHub.

[52] See the one found on page 2 of the present DMP.

3.2 Making the Data Accessible

All metadata in the repositories already mentioned are made publicly available and freely accessible under the CC0 licence.[53] They are expressed in common and consolidated standard formats to facilitate their reading, reuse and so the access to the data themselves, which are licenced under CC-BY,[54] an open specification with an active community, to also reach a broader audience.

All the data and metadata will also be accessible by default on the DHARMA-base. Access to metadata and data files is provided over standard communication protocols such as HTTP while in due course it may become interesting to deploy a REST-API.[55]

In general, we intend to impose as little restriction on the accessibility of data in the DHARMA-base as possible, although it will not be possible to offer all of the data entirely free of restrictions for reading or reuse.[56]

If it is necessary to impose a temporary embargo on some scholarly publications and research data, this can be directly handled both in Zenodo and in HAL-SHS. In this case, the repositories themselves will restrict access until the end of the embargo period determined by the users when depositing the digital objects. Nevertheless, even if an embargo is imposed on the content, the metadata will still be accessible as well as retrievable.

The digital editions of inscriptions or manuscripts will be accessible on GitHub with the possibility to download them, from the beginning of the project. Official releases will be deposited on Zenodo, once the work has been validated by the researchers working on the repositories and in agreement with the project's XML-TEI Data Manager. Note that not all project members will be granted the same access rights. Three levels of rights have been established and will be applied to all the tools of the project.

- "Observer rights" to read a file without being able to delete or modify it
- "Member rights" to read, write, modify and delete
- "Administrative rights" comprising all "member rights" but additionally include the ability to change configurations and settings (e.g., deleting repositories or changing their names)

The access rights will be set so that everyone in the project will have access as "observer" to all the working files of the whole team. "Member rights" will be given to the group involved in editing a given corpus, while "administrative rights" will be given to the PIs, designated coordinators of certain corpora, and the Visual Data Manager or XML-TEI Data Manager depending on the tool.

3.3 Making the Data Interoperable

The DHARMA project will be using standard metadata schemas and protocols — all intended to optimize the interoperability of its data. While several possibilities are currently under study to reuse existing vocabularies and ontologies, it presently seems most likely that we will be

[53] https://creativecommons.org/publicdomain/zero/1.0/deed.fr
[54] https://creativecommons.org/licenses/by/2.0/
[55] REST is an open, free and universal protocol to retrieve information on the web.
[56] Archeologists involved in the project have brought to light possible restrictions on some data to protect the archeological sites from possible looters. Object descriptions could be less detailed and more general than for other datasets and the access to location information could be restricted to team members and *bona fide* scholars. To take decisions on the matter, we await the first results of the archaeological work to be carried out as part of DHARMA.

making our own ontology in close alignment with modelisation established by our Visual Data Manager in another project.[57] The ontology will structure the metadata following the model provided by the CIDOC Conceptual Reference Model dedicated to Cultural Heritage Data.[58]

To manage vocabularies, we will use the tool OpenTheso,[59] which allows us to reuse any existing thesaurus or to create our own. This open-source software helps to manage vocabularies, to build the terminological hierarchy and manage synonyms as well as corresponding terms in the various Asian and European languages relevant to DHARMA. Each term has its own persistent ARK identifier[60] and can be aligned on external vocabularies.[61]

Some external controlled vocabularies will be used to describe our data and to complete our DCAT metadata schema, e.g. Creatives Commons licences, funders and grants from the funder registry maintained by the CrossRef Organization.[62] Each referenced external piece of metadata will be qualified by a resolvable URL.

3.4 Making the Data Reusable

All the datasets will be shared as openly as possible to encourage reuse of our work. To achieve this, all the files produced will have a licence or rights statements declared in their metadata. The same will be redeclared in the metadata forms included in our repositories. The majority of the work will be licensed under the Creative Commons CC-BY and the metadata as CC0.[63] When possible those licences will be recorded with their URI. Depending on the rights given by third parties like museums or archeological services, we may need to use more restrictive licences. Some transformation and code files may require a software licence, such as the MIT License or GNU General Public License.[64]

The authors and the provenance of the data will always be explained in the metadata. If the data are reused from another work or project, the information related to authors, project and provenance will also be given with a description of what has been modified and by whom in the context of the DHARMA project. By uploading content into our repositories, ownership does not shift and the property rights are not automatically owned by the owner of the repository.

---

[57] An ontology has been developed by Adeline Levivier, our current Visual Data Manager, in the context of her PhD dissertation on Greek epigraphy, for the E-STAMPAGES project which aimed to make Greek estampages findable and accessible online, in order to study estampages, inscriptions and artefacts. See https://www.e-stampages.eu/.

[58] http://www.cidoc-crm.org/

[59] OpenTheso (https://github.com/miledrousset/opentheso) is developed by Miled Rousset for the network Frantiq. Its purpose is to manage archaeological, historical and epigraphical vocabularies. See also https://thesaurus.mom.fr/opentheso/index.xhtml.

[60] https://n2t.net/e/ark_ids.html

[61] Such as those of the Getty Museum (https://www.getty.edu/research/tools/vocabularies/), the EAGLE network (https://www.eagle-network.eu/resources/vocabularies/), etc.

[62] Crossref (https://www.crossref.org/) is an official DOI Registration Agency of the International DOI Foundation.

[63] We choose these licences because their use is widespread, flexible and easy. They are readable both by humans and machines allowing researchers and computers to know what they can and cannot do with given data.

[64] The MIT License (https://opensource.org/licenses/MIT) is a permissive free software license. The GNU General Public License (https://www.gnu.org/licenses/gpl-3.0.fr.html) is a free, “copyleft” license for software and other kinds of works.

4. Allocation of Financial Resources

4.1 Provisional Costs

ShareDocs and Huma-Num Box are free services offered by Huma-Num, while some of the services offered by GitHub and Zotero need to be paid for. “Professional”, i.e. for-payment, GitHub accounts are not required for any team-member. It is possible to request an “educational” account to gain cost-free access to some services that would otherwise need to be paid for. But at this stage, we do not foresee such an account to be required for any member of the project.

Zotero offers free storage space of 300 MB, which may not be enough for the needs of the project. For the first year, we will use the free storage space and when necessary upgrade it to 2GB, 6GB or unlimited storage space in year 2 or a later year of the project. These options cost respectively 20 US$, 60 US$ and 120 US$ per year, being 17,91 €, 53,72 € and 108,63 €. So, at most it will cost 600 US$, being 548,07 €, for the remainder of the project.

Some proprietary tools are necessary. When possible, the host institutions will provide the licences to reduce the cost covered by the project budget. Nevertheless, the GA sanctions a budget for consumables: 3.000 € at the CEIAS-CNRS, 7.500 € at the EFEO and 1.200 € at the UNO.[65]

An amount of 100.000 € has been budgeted to develop the DHARMA-base or outsource code writing as well as to buy a namespace and server space.[66] A further 50.000 € have been provisioned to update and maintain the DHARMA-base. The actual costs will become clearer at a later stage of the project.

4.2 Costs to Make the Data FAIR

The repositories Zenodo, HAL-SHS and Didómena make it possible to deposit scholarly publications as well as research data in Open Access free of charge. Zenodo and Didómena provide DOIs, while HAL-SHS does not.

Buying rights from publishers may be necessary to provide Open Access to some of the results. Those costs are eligible for reimbursement on the project budget during the duration of the project under conditions defined in the GA.

4.3 Role Distribution

The tasks and responsibilities are shared between members of the project.

The researchers are responsible of their metadata during the project as well as for scholarly quality control. We plan to appoint Data Captains at corpus level to help the PIs carry out internal reviews. The role of the PIs is to supervise all creation and modification of data and to validate their completeness and integrity from the scholarly point of view. The Visual Data Manager and the XML-TEI Data Manager are responsible for quality control of the metadata and formal validation of all the datasets.

---

[65] Those should cover the Oxygen editor licence that cost around 99 US$, being 89 €, per year and some software applications like Autocad map, Covadis, Agisoft Metashape, FileMakerPro and Adobe Creator Suite. Although no such budget line was foreseen by UBER, this host institution will provide Oxygen licences and other needed software without charging costs to the project.

[66] Around 30 € per year are to be expected for a basic package at a good server farm.

The researchers are responsible for their datasets in the preparatory stages before their addition into any tools chosen for the project. They have been advised to maintain their own backups according to the 3-2-1 law.[67] Help will be provided to set up a backup protocol for those without one to ensure the safety of their data. In addition, since the data must be secured for at least 5 years after the end of the project, the members are required to preserve and curate the data, e.g. update the persistent identifiers, publish a correction of datasets if they disappear in whole or in part or if any part becomes corrupt.

5. Data Security

5.1 Data Storage

The ShareDocs service allows us to store the current data and documentation insofar as we are dealing with PDFs and image files. It is based on the FileRun application which supports standard security protocols (XSS and SQL injection protection, HTTPS/SSL). Each user of the workgroup DHARMA has his/her own account and access to all or some of the folders and files. The access rights can be restricted (reading/modifying). Huma-Num makes regular backups of the repositories, through the IN2P3's Computing Center based in Lyon.

Transfers of data to GitHub are encrypted with standard security protocols (SSH and HTTPS/SSL). The company has its own data centers. The content itself is encrypted if GitHub requires a third-party storage provider.

Zotero is also aligned on the current best practices (HTTPS/SSL). The storage itself is located in the Amazon Cloud, in the us-east-1 AWS region in Virginia. Regular automated backups are made to protect against accidental loss of data. These backups are intended for disaster recovery only and they may be retained for up to 6 months.

Huma-Num Box is a NAS service that works with an SFTP client. Its nine servers allow us to store the heavier data (such as RAW images) to which we don't require access on a daily basis. Storage in Huma-Num Box is considered as intermediate archiving.

Didómena, HAL-SHS, Zenodo will be used for long-term storage and have their own security policies, operationalized by the hosting infrastructures, i.e., the Penn State Data Center Services for Didómena, the IN2P3's Computing Center and the CNRS for HAL-SHS and the CERN for Zenodo.

5.2 Data Transfer

Our data transfer will be handled with tools based on web services using the HTTP protocol. Non-sensitive files — the vast majority of our data — will be directly uploaded into the relevant GitHub repository or ShareDocs folders, depending on the type of file. Sensitive data will be exchanged exclusively in restricted-access ShareDocs folders.

[67] The 3-2-1 backup law supposes that each file exists in three versions: the original and two duplicates kept in two different locations.

6. Ethical and Legal Aspects

All datasets handled in our repositories will respect the intellectual property rights. The copyright ownership of the data follows national as well as European legislation. Each dataset generated will be associated with a licence with appropriate attribution of copyright ownership in CC-BY, while offering the possibility to third parties to reuse them.

Any member of the project has the possibility to ask for an adjustment to the CC-BY licence if it seems necessary. For those cases, an agreement must be reached with all the members involved with the data generation and the decision must be approved by the PIs.

Data identified as sensitive may not be shared. The management of private and personal data are the responsibility of the producers. The members handling those datasets determine who else will be given access. Note that in this case, the data won't be given in open-access and won't be reusable.

All project participants have been warned that use of tools like GitHub means that some data will be accessible to the public even while they are being worked on. However, only members of the project are allowed to upload content to any of the project's repositories. Team members have a right to anonymity and may express to the PIs their desire to remain anonymous.

When submitting a data file in any of the repositories, the submitter has to verify that the dataset doesn't contain any personal data. The submitter also has to make sure that by submitting metadata and data, he is acting with the consent of all persons involved since any such file will contain a declaration of the name and surname of the participating researchers. Inversely, the submitter must check that he does not violate anyone's rights by not declaring that person's name inside the file. To the extent that they need to use proprietary software, the team members have been informed about the moral and legal imperative to work under proper licences and are actively encouraged to work under licences purchased by the project when creating and editing their data files.

If there are any legal or ethical issues that need to be discussed more thoroughly, this will be done in the Ethical report, work package 2, due April 30th, 2020.

7. Acknowledgement & Disclaimer

The DHARMA project has received funding from the European Commission under the Horizon 2020 programme, Grant Agreement no. 809994.

The opinions expressed in this document reflect only the project consortium's views and are in no way representative of the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information this document contains.

The consortium has tried as much as possible for the content to be accurate, consistent and lawful. However, neither the project consortium nor the individual members who actively participated in the conception of this Data Management Plan accept any liability for unforeseen consequences of reusing its content.