Project Number: 101087269

TALOS AI for SSH
University of Crete

Funded by the
European Union

# D1.2

# DATA MANAGEMENT PLAN

Date: 31 August 2023

# TALOS AI for SSH - DMP

# Data Management Plan – D1.2 – 1st Version

## Project

| | | | |
|---|---|---|---|
| **Project number:** | 101087269 | | |
| **Project name:** | TALOS-Artificial Intelligence for Humanities and Social Sciences | | |
| **Project acronym:** | TALOS-AI4SSH | | |
| **Call:** | HORIZON-WIDERA-2022-TALENTS-01 | | |
| **Topic:** | HORIZON-WIDERA-2022-TALENTS-01-01 | | |
| **Type of action:** | HORIZON-CSA | **Service:** | REA/C/03 |
| **Project starting date:** | 1 March 2023 | **Project duration:** | 60 months |

| | | |
|---|---|---|
| **ERA Chair Holder:** | Prof Christophe Roche | christophe.roche@uoc.gr |
| **Co-ordinator:** | Prof Melina Tamiolaki | tamiolaki@uoc.gr |
| **Project Manager:** | Dr Maria Papadopoulou | maria.papadopoulou@uoc.gr |

| | |
|---|---|
| **Web site:** | https://talos-ai4ssh.uoc.gr/ |

## Deliverable / Milestone / Report

| | | | |
|---|---|---|---|
| **Deliverable / Milestone name:** | Data Management Plan | | |
| **Deliverable / Milestone number:** | D 1.2 | **Work Package** | WP1 - Management |
| **Due date (month):** | Month 6 – August 2023 | | |
| **Dissemination level:** | PU (Public) | **Type** | Report |
| **Status:** | 1st Version | | |
| **Authors (organisation, email):** | C. Roche (CR) christophe.roche@uoc.gr<br>M. Papadopoulou (MP) maria.papadopoulou@uoc.gr | | |
| **Contributors (organisation, email):** | | | |
| **Reviewers (organisation, email):** | | | |

## History of Changes

| Revision | Date | Author | Organisation | Description |
|---|---|---|---|---|
| V 0.1 | 16/06/2023 | CR, MP | UoC | 1st Draft |
| V 0.2 | 01/08/2023 | CR | UoC | Revised Draft |
| V 0.3 | 04/08/2023 | MP, CR | UoC | Revised Draft |
| V 1.0 | 26/08/2023 | MP, CR | UoC | 1st Version |

## Contents



# 1. Definitions and acronyms

| | |
|---|---|
| AB | Advisory Board |
| AI | Artificial Intelligence |
| BA | Bachelor |
| CC | Creative Commons |
| CH | Cultural Heritage |

| | |
|---|---|
| Dereferencing | The act of retrieving a representation of a resource identified by a URI |
| DH | Digital Humanities |
| DL | Deep Learning |
| DMP | Data Management Plan |
| DS | Data Set |
| EC | European Commission |
| EOSC | European Open Science Cloud |
| FAIR | Findable, Accessible, Interoperable, Reusable |
| GA | Grant Agreement |
| GDPR | General Data Protection Regulation |
| HAI | Hybrid Artificial Intelligence |
| ISO | International Organization for Standardization |
| LOD | Linked and Open Data |
| MA | Master |
| Metadata | A description of data |
| ML | Machine Learning |
| MOOC | Massive Open Online Course |
| NLP | Natural Language Processing |
| Open Access | Access that is free to all and free of any restrictions |
| Open Data | Data that can be freely used, shared and built on by anyone for any purpose |
| PDECA | Plan for Dissemination, Exploitation and Communication Activities |
| Persistent URI | A long-lasting reference to a web resource |
| PMB | Project Management Board |
| Repository | A location in which data are stored or managed |
| RT | Research Topic |
| SAI | Symbolic Artificial Intelligence |
| SSH | Social Sciences & Humanities |
| SW | Semantic Web |
| UoC | University of Crete |
| W3C | World Wide Web Consortium consortium |
| WP | Work Package |

## 3. Executive summary

This deliverable introduces the first version of the Data Management Plan (DMP) for the TALOS AI for SSH project, before the thrust of its research activities and therefore before generating, collecting and reusing any set of data.

The TALOS DMP outlines the ways in which data are generated, collected, handled, and processed throughout the life of the project and after completion of the project. It aims to ensure data availability, sustainability and reuse for further purposes and applications. The DMP will be updated and adjusted regularly in line with the progress of the project.

All TALOS data will comply with the FAIR (Findable, Accessible, Interoperable, Reusable) principles and with the European policies on standards and best practices, as well as with the reuse ethos of the W3C (World Wide Web Consortium consortium).

## 4. Introduction

The TALOS project in AI for SSH aims to create and manage a new centre of excellence in explainable AI for SSH (Humanities and Social Sciences). The TALOS Center interconnects the three departments of the Faculty of Philosophy at UoC Rethymnon (History and Archaeology, Philosophy and Social Sciences) under a common research and teaching agenda.

To reach the research objectives, six Research Topics (RTs) were set up. Four of them called Horizontal RTs, from RT 1 to RT 4, are directly linked to SSH: semantic annotation of texts, preservation and dissemination of Cultural Heritage, corpora analysis, and digitalisation of education. These four Horizontal RTs re-use data, such as the Perseus Digital Library, and produce data such as annotated resources, ontologies and terminologies. Two Transversal RTs linked the Horizontal RTs by providing common resources such as software (RT 5) and standards (RT 6)(Fig. 1).

The TALOS Center teaching activities aim to disseminate AI literacy & ethics to different target audiences through innovative modules: curricula, courses, seminars, course materials, and MOOCs. For the TALOS' dissemination, exploitation and communication policy please refer to the Plan for Dissemination, Exploitation and Communication Activities (PDECA).

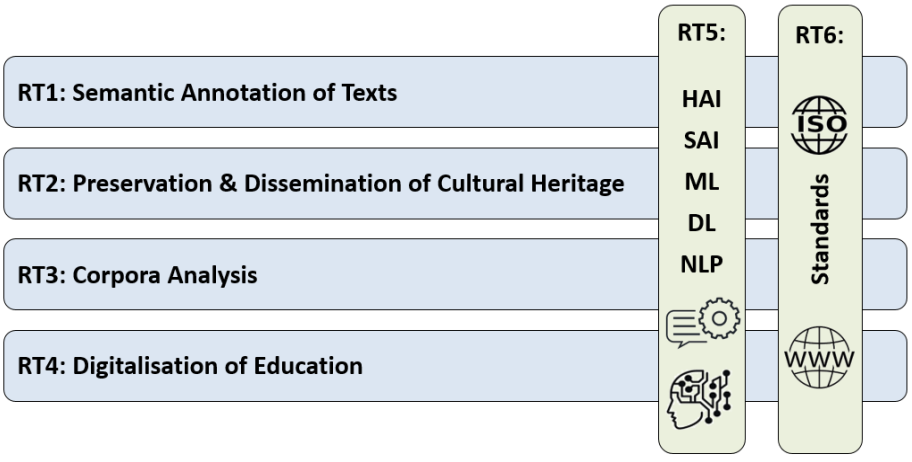


Fig. 1 The TALOS Research Topics

# 5. Data summary

Existing data are reused and new data are to be produced by the Horizontal Research Topics of the WP2 (RT1 to RT4). The tables below describe, for each RT, the data which are reused and the data that will be produced, as well as their origin, format, and their targeted users.

**The RTs start at Month 7 (September 2023), the information about data is indicative.**

### RT1: Semantic Annotation of texts

Reused data

| Dataset name | Origin | Link to the RT | Format |
|---|---|---|---|
| Perseus Digital Library | TUFTS University | - Corpus to be annotated<br>- Input data for ontology and terminology building | XML TEI |

Generated data

| Dataset name | Origin | Link to the RT | Format | Targeted users |
|---|---|---|---|---|
| Set of stand-off annotations | Corpora | Set of stand-off annotations | RDF | SW, AI, NLP, SSH and DH communities |
| Domain Ontology | From input data and experts | Used for annotating corpora | RDF OWL | SW, AI, NLP, SSH and DH communities |
| Domain Terminology | From input data and experts | Used for annotating corpora | RDF | SW, AI, NLP, SSH and DH communities |
| Electronic Dictionary | Terminology, Ontology and experts | Dictionary of the domain | HTML | SSH and DH communities |

### RT2: Preservation and dissemination of Cultural Heritage

Re-used data

| Dataset name | Origin | Link to the RT | Format |
|---|---|---|---|
| DAMOS | Database of Mycenaean at Oslo | Creating a digital resource | |
| Epigraphic Database Heidelberg (EDH) | Epigraphic Database Heidelberg | Epigraphic Database | XML RDF JSON |
| Inscriptiones Graecae (IG) | Berlin-Brandenburg Academy of Sciences and Humanities | Epigraphic Database | physical epigraphic objects |

Generated data

| Dataset name | Origin | Link to the RT | Format | Targeted users |
|---|---|---|---|---|
| Images of artefacts | Databases | Contribution to a digital resource | RDF | DH communities |
| Textual materials | Databases | Contribution to a digital resource | RDF TEI | DH communities |

**RT3: Corpora analysis**

Re-used data

| Dataset name | Origin | Link to the RT | Format |
|---|---|---|---|
| Greek literary corpus | | Used for NLP analysis | TXT |
| Named Entity Dataset | | Used for annotating | |

Generated data

| Dataset name | Origin | Link to the RT | Format | Targeted users |
|---|---|---|---|---|
| 19th century literary corpus in .txt format | | Result of the RT | TXT | DH communities |
| Annotated corpus | | Result of the RT | TXT | DH communities |

**RT4: Digitalisation of education**

Re-used data

| Dataset name | Origin | Link to the RT | Format |
|---|---|---|---|
| Greek Curricula for L1 | Greek Institute of Educational Policy | Input data for ontology and terminology building | PDF TXT |

Generated data

| Dataset name | Origin | Link to the RT | Format | Targeted users |
|---|---|---|---|---|
| Terminology | From input data and experts | Used for annotating corpora | RDF | Education communities |
| Ontology | From input data and experts | Used for annotating corpora | RDF | Education communities |
| Annotated Curricula | From input data and experts | Result of the RT | RDF | Education communities |

# 6. FAIR data

## 6.1 Making data findable, including provisions for metadata

Rich metadata, including keywords, will be associated with the TALOS data to make their discovery easier. As one of the main objectives of TALOS is to produce open-access data for the semantic web, W3C standards and principles will be followed.

## 6.2 Making data accessible

Repository:

The main repository for TALOS data will be hosted by the University of Crete (UoC) which will ensure sustainability. The data will also be deposited in trusted repositories (e.g., Zenodo https://zenodo.org/), according to the EC guidelines (https://open-research-europe.ec.europa.eu/for-authors/data-guidelines). Where applicable, data will be stored in open archives handling persistent identifiers.

Data:

All data produced by TALOS RTs will be in open access (there is no need for a data access committee to evaluate/approve access requests to personal/sensitive data). No embargo period will be applicable. There will be no restrictions on their use. Access will be provided via a web browser dedicated to the TALOS data. This will enable documentation and downloading functionalities, as well as a SPARQL endpoint for RDF data queries.

Metadata:

The data and metadata will be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement.

The data and metadata will remain available and findable for 5 years after the end of the project. Using standard formats, no specific software is needed for accessing the data.

## 6.3 Making data interoperable

Community standards: TALOS will use XML-oriented standards, e.g. TEI, including W3C Standards, such as RDF/RDFS, DC, SKOS, FOAF, OWL. The W3C community standards and ISO principles will be followed for building the datasets.

Mappings: The RTs will produce specific ontologies and terminologies (vocabularies) in RDF formats. Where applicable, they will be linked and mapped to more commonly used ontologies such as CIDOC-CRM. All of them will be openly published to allow reusing, refining or extending them.

The LOD and FAIR principles for interoperability are going to be followed for all produced datasets.

The data will include, where applicable, qualified references to other data from the TALOS project, or other open datasets from previous research.

## 6.4 Increase data re-use

Documentation about the provenance of the data, methodology, used standards, statistics, and specific vocabularies will be provided to facilitate data reuse.

All the datasets will be made freely available in the public domain to permit the widest reuse by third parties during and after the end of the project.

Data quality assurance processes: Using W3C and ISO principles for dataset building guarantees data quality. Further information on quality assurance processes will be included in the Quality and Risk Management Plan (D1.3/D3).

# 7. Other research outputs

### RT5: HAI-NLP Software for SSH

The RT5 is dedicated to AI software used and implemented by TALOS to carry out the RTs. The software implemented by TALOS will be open source and deposited in GitHub and Sourceforge.

At this point, it is not yet possible to exactly identify which software will be needed and which will be developed. The list below is indicative.

Re-used software

| Software | Description | RT | Origin & Format |
|---|---|---|---|
| Protégé | Ontology building environment | All RTs | Stanford University<br>Open source |
| Colab | Access to computing resources for ML | RT3 | Google product |
| Antconc | Corpus analysis toolkit | RT1, RT4 | Waseda University, Japan<br>© Laurence Anthony, Wasea University |
| TermoStat | Corpus analysis term extraction tool | RT1, RT4 | Free for research<br>© P. Drouin, Montréal University, Canada |
| TEDI | Ontoterminology building environment | RT1, RT4 | Free for research<br>© C. Roche, UoC |

Software to implement for TALOS

| Software | Description | RT | Format | Targeted users |
|---|---|---|---|---|
| TALOS KGE | Online Knowledge Graph Editor | RT1, RT2, RT4 | Open source | SW, AI, NLP, SSH and DH communities |
| TALOS Dictionary | Online platform for writing and publishing dictionaries | RT1, RT2, RT3 | Open source | SW, AI, NLP, SSH and DH communities |
| TALOS Semantic Annotation | Semantic Annotation of Corpora | RT1, RT4 | Open source | SW, AI, NLP, SSH and DH communities |

### RT6: Standards for SSH

The RT 6 focuses on standards for AI for SSH including knowledge graphs, ontologies, thesauri, developed within the framework of the RTs with the aim to further their dissemination, sharing and exploitation in future projects. These standards correspond to terminologies, vocabularies and ontologies re-used and developed in the Horizontal RTs (RT1 to RT4). They are listed in the previous chapter (See “5. Data summary”).

## 8. Allocation of resources

Datasets are produced in the RTs of the WP2 (Research Management) whose allocated person-months are the most important of the whole TALOS budget. At this point, it is not possible to assess the exact allocation of resources for making data FAIR. Data hosted in repositories at UoC will be free of charge. The cost of archiving data outside UoC will be assessed as soon as external storage is identified.

The Co-ordinator and the Project Manager are responsible for data management. The data and metadata will remain available and findable for 5 years after the end of the project.

## 9. Data security

There are no sensitive data in TALOS. TALOS data are safely stored in repositories handled by UoC for long-term preservation and curation. A periodic data-backup system has been set up.

## 10. Ethics

TALOS does not handle any personal data. There are no legal or ethical issues raised by sharing TALOS data.

## 11. Next step

The DMP will be updated and adjusted regularly in line with the progress of the project. The DMP-Update 1 (D1.4/D4) is due on August 31st, 2025 and the final version of the DMP is due on November 30th, 2027 (D1.7/D7).

_____*End of the Report*