



**TALOS AI for SSH**  
University of Crete



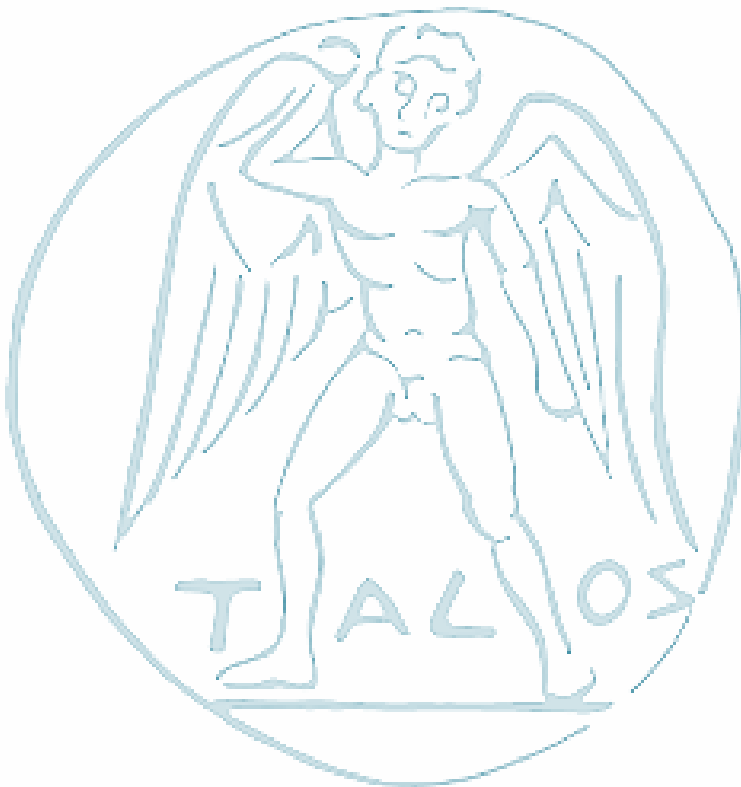
Funded by the  
European Union

# D1.4

# DATA

# MANAGEMENT

# PLAN – Update 1



Date: 31 August 2025

# TALOS AI for SSH - DMP

## Data Management Plan – D1.4 – 1<sup>st</sup> Update

### Project

<b>Project number:</b>	101087269		
<b>Project name:</b>	TALOS-Artificial Intelligence for Humanities and Social Sciences		
<b>Project acronym:</b>	TALOS-AI4SSH		
<b>Call:</b>	HORIZON-WIDERA-2022-TALENTS-01		
<b>Topic:</b>	HORIZON-WIDERA-2022-TALENTS-01-01		
<b>Type of action:</b>	HORIZON-CSA	<b>Service:</b>	REA/C/03
<b>Project starting date:</b>	1 March 2023	<b>Project duration:</b>	60 months
<b>ERA Chair Holder:</b>	Prof Christophe Roche	christophe.roche@uoc.gr	
<b>Co-ordinator:</b>	Prof Melina Tamiolaki	tamiolaki@uoc.gr	
<b>Project Manager:</b>	Dr Maria Papadopoulou	maria.papadopoulou@uoc.gr	
<b>Web site:</b>	<a href="https://talos-ai4ssh.uoc.gr/">https://talos-ai4ssh.uoc.gr/</a>		

### Deliverable / Milestone / Report

<b>Deliverable / Milestone name:</b>	Data Management Plan		
<b>Deliverable / Milestone number:</b>	D 1.4	<b>Work Package</b>	WP1 - Management
<b>Due date (month):</b>	Month 30 – 31 <sup>st</sup> August 2025		
<b>Dissemination level:</b>	PU (Public)	<b>Type</b>	Report
<b>Status:</b>	1 <sup>st</sup> update of the DMP		
<b>Authors (organisation, email):</b>	M. Papadopoulou (MP) <a href="mailto:maria.papadopoulou@uoc.gr">maria.papadopoulou@uoc.gr</a> C. Roche (CR) <a href="mailto:christophe.roche@uoc.gr">christophe.roche@uoc.gr</a>		
<b>Contributors (organisation, email):</b>			
<b>Reviewers (organisation, email):</b>			

### History of Changes

Revision	Date	Author	Organisation	Description
V 0.1	01/06/2025	MP	UoC	1st Draft
V 0.2	08/08/2025	CR; MP	UoC	Revised Draft
V 0.3	11/08/2025	MP; CR	UoC	Revised Draft
V 0.4	22/08/2025	MP; CR	UoC	Revised Draft
V1.0	31/08/2025	EMT	UoC	Final Draft

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

## Contents

<b>1. Definitions and acronyms.....</b>	<b>3</b>
<b>2. International Organization for Standardization.....</b>	<b>4</b>
<b>3. Executive summary.....</b>	<b>5</b>
<b>4. Introduction.....</b>	<b>5</b>
<b>5. Data summary.....</b>	<b>6</b>
RT1: Semantic Annotation of texts.....	6
RT2: Preservation and Dissemination of Cultural Heritage.....	7
RT3: Corpora analysis.....	7
RT4: Digitalisation of education.....	8
<b>6. FAIR data.....</b>	<b>8</b>
6.1 Making data findable, including provisions for metadata.....	8
6.2 Making data accessible.....	8
6.3 Making data interoperable.....	9
6.4 Increase data re-use.....	9
<b>7. Other research outputs.....</b>	<b>9</b>
RT5: HAI-NLP Software for SSH.....	9
RT6: Standards for SSH.....	10
<b>8. Allocation of resources.....</b>	<b>11</b>
<b>9. Data security.....</b>	<b>11</b>
<b>10. Ethics.....</b>	<b>11</b>
<b>11. Next step.....</b>	<b>11</b>

## 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 update of the Data Management Plan (DMP) for the TALOS AI for SSH project, after the thrust of its research activities in all RTs.

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).

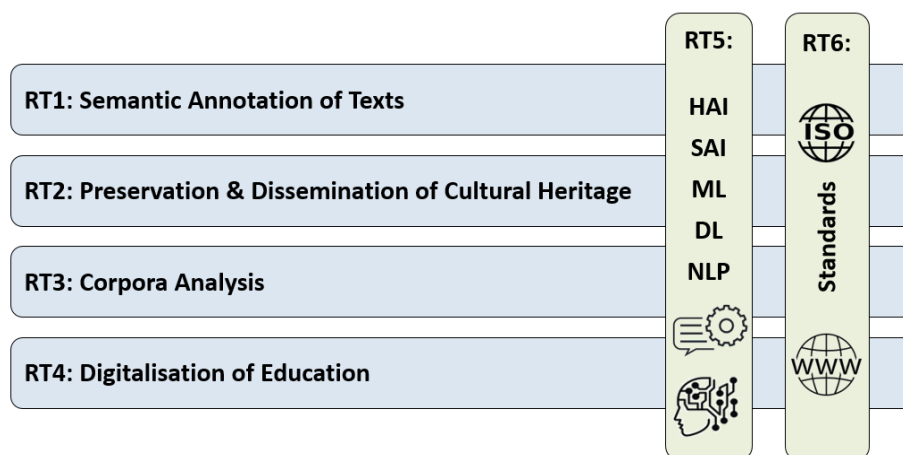


Fig. 1 The TALOS Research Topics

## 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 started 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 - Input data for ontology and terminology building	XML TEI

#### Generated data

##### Note:

All new versions of TALOS datasets will be downloadable in Zenodo by April 31, 2026.

All datasets are also available via TALOS website <https://talos-ai4ssh.uoc.gr/>>Resources>Open Datasets.

Dataset name	Origin	Link to the RT	Format	Targeted users
Set of stand-off annotations	Corpora	Set of stand-off annotations.  (Produced by TALOS PhD cand. Rachel Milio under the supervision of Prof. Roche & Dr. Papadopoulou).	RDF	SW, AI, NLP, SSH and DH communities
Attic Law & Terminology	From input data and experts	Used for annotating corpora.  Domain Ontologies.  (Produced by TALOS PhD cand. Rachel Milio under the supervision of Prof. Roche & Dr. Papadopoulou).  <a href="#">link</a>	RDF SKOS OWL OntoLex OTV	SW, AI, NLP, SSH and DH communities

Attic Law Electronic Dictionaries	Terminology, Ontology and experts	Dictionaries of the domain (produced by TALOS PhD cand. Rachel Milio under the supervision of Prof. Roche & Dr. Papadopoulou).	HTML	SSH and DH communities
Göbekli Tepe dataset	Terminology, Ontology and experts	Open Datasets <a href="#">link</a> (produced by TALOS Research Assistants, Rafail Giannadakis & Tonia Lourentzaki, under the supervision of Prof. Roche & Dr. Papadopoulou).	RDF SKOS OWL OntoLex OTV	SSH and DH communities
Hellenistic events dataset	Terminology, Ontology and experts		RDF SKOS OWL OntoLex OTV	SSH and DH communities
ALyRA dataset	Terminology, Ontology and experts		RDF SKOS OWL OntoLex OTV	SSH and DH communities
Philosophers dataset	Terminology, Ontology and experts		RDF SKOS OWL OntoLex OTV	SSH and DH communities

## RT2: Preservation and Dissemination of Cultural Heritage

### Re-used data

The PhD for RT2 was hired in March 2025. She is currently working on building a corpus of images of wall paintings at Knossos during the Mycenaean period. A first version of the ontoterminology of Mycenaean Wall Painting is available [here](#). The first official version will be uploaded in Zenodo before April 30, 2026.

Dataset name	Origin	Link to the RT	Format
Wall Painting	The images are not publicly accessible at this point, as some licences need to be obtained.	Creating an ontology	jpeg, png

### Generated data

Dataset name	Origin	Link to the RT	Format	Targeted users
--------------	--------	----------------	--------	----------------

Digital Archeology	Databases/Internet images of artefacts, stored <a href="#">here</a> (collected by RT2 PhD Cand. Melissa Bergoffen). ernet	Contribution to a digital resource	RDF	DH communities
--------------------	---	------------------------------------	-----	----------------

### RT3: Corpora analysis

#### Re-used data

Dataset name	Origin	Link to the RT	Format
Anemi digital library <a href="https://anemi.lib.uoc.gr/">https://anemi.lib.uoc.gr/</a>	University of Crete Library	Corpus building source	PDF
Psifiothiki <a href="https://digital.lib.auth.gr/?ln=el">https://digital.lib.auth.gr/?ln=el</a>	Aristotle University of Thessaloniki Library	Corpus building source	PDF
Project Gutenberg <a href="https://www.gutenberg.org/">https://www.gutenberg.org/</a>	Project Gutenberg Literary Archive Foundation	Corpus building source	TXT
Google Books <a href="https://books.google.com/">https://books.google.com/</a>	Google LLC	Corpus building source	PDF
Ellinikos politismos <a href="https://users.sch.gr/ipap/Ellinikos%20Politismos/eisodos.htm">https://users.sch.gr/ipap/Ellinikos%20Politismos/eisodos.htm</a>	Greek School Network (GSN)	Corpus building source	HTML, TXT
El_core_news_lg <a href="https://spacy.io/models/el">https://spacy.io/models/el</a>	spaCy NLP library	Used for fine-tuning NER on 19th-century Greek fiction	.py
GreekBERT (pretrained corpora of bert-base-greek-uncased-v1) <a href="https://huggingface.co/nlpauab/bert-">https://huggingface.co/nlpauab/bert-</a>	Hugging Face	[Tokenizer] Used for fine-tuning NER on 19th-century Greek fiction	.py

<a href="#">base-greek-uncased-v1</a>			
---------------------------------------	--	--	--

Generated data

Dataset name	Origin	Link to the RT	Format	Targeted users
Documentation of the 19th-century corpus of Ionian Fiction	From input data (Produced by TALOS PhD cand. Foteini Koidaki under the supervision of Prof. Chatzikyriakidis)	Used for recording and documenting the corpus-building process (location of resources, digitization); includes derived visualizations (e.g. graphs, charts, maps) produced from the tabular dataset.  <a href="#">Link</a>	XLSX	Greek Literary Studies and DH communities
NER Training Data for Geoparsing the 19th-century Greek Prose Fiction	From digitized texts of the Ionian Prose Fiction Corpus (Produced by TALOS PhD cand. Foteini Koidaki under the supervision of Prof. Chatzikyriakidis)	Supporting the fine-tuning of NER models for Location Extraction/ within the TALOS (RT3) project corpus; includes accompanying training scripts for BERT and spaCy.  <a href="#">Link</a>	TXT, JSON, CoNLL2003	AI, NLP, SSH and DH communities
Gazetteer-based Patterns for NER (Greek Toponyms)	Compiled from Wikipedia entries of Greek toponyms (Produced by TALOS PhD cand. Foteini Koidaki under the supervision of Prof. Chatzikyriakidis).	Supporting the rule-based extraction of location entities using spaCy's EntityRuler in the TALOS (RT3) project corpus.  <a href="#">Link</a>	TXT (patterns), PY (script)	NLP, DH, SSH communities
Transkribus HTR Model "19th Century Greek 8.0"	Trained on digitized texts of the Ionian Prose Fiction Corpus (Produced by TALOS PhD cand. Foteini Koidaki under the	Openly available model at <a href="#">Transkribus Model Repository</a> .	HTR Model (PyLai architecture)	AI, NLP, SSH and DH communities

	supervision of Prof. Chatzikyriakidis).			
--	---	--	--	--

#### 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 Open source
Colab	Access to computing resources for ML	RT3	Google product
Antconc	Corpus analysis toolkit	RT1, RT4	Waseda University, Japan © Laurence Anthony, Wasea University

TermoStat	Corpus analysis term extraction tool	RT1, RT4	Free for research © P. Drouin, Montréal University, Canada
TEDI	Ontoterminology building environment	RT1, RT4	Free for research © C. Roche, UoC

Software to implement for TALOS (openly accessible via TALOS website <https://talos-ai4ssh.uoc.gr/>Resources>Software>)

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
TALOS RDF Viewer	RDF Graph Viewer	RT1, RT2, RT4	Open source	SW, AI, NLP, SSH and DH communities
TALOS Text Analyser	Multilingual text file analyser, NER, Pattern extraction	RTs	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, and 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”).

#### Re-used Standards

Standards	Description	RT	Origin & FormatTE
TEI	TEI (Text Encoding Initiative): standardized framework for representing and encoding texts in digital form	RT1, RT4	Open source
RDF-RDFS	W3C Formats for Linked and Open Data, and Semantic Web	RT1, RT2, RT4	W3C Open source
SKOS	RDF Language dedicated to Simple Knowledge Organisation System	RT1, RT2, RT4	W3C Open source

OWL	Ontology Web Language. A RDF language dedicated to Ontology Representation	RT1, RT2, RT4	W3C Open source
CIDOC-CRM	international standard for the controlled exchange of cultural heritage information	RT2	ISO 21127:2023 Open source
OntoLex-Lemon	vocabulary for publishing lexical data as a RDF Knowledge Graph	RT1	W3C Open source

#### Standard to implement for TALOS

Software	Description	RT	Format	Targeted users
<a href="#">OTV</a>	OntoTerminology Vocabulary: RDF vocabulary dedicated to representing ontoterminology in RDF format	RT1, RT2, RT4	RDF Open source	Humanists

## 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.

The datasets created by TALOS as of August 31, 2024 are also stored in Zenodo and are accessible via TALOS website: <https://talos-ai4ssh.uoc.gr/resources/open-datasets/>

## 10. Ethics

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



## 11. Next step

New versions of the datasets are produced periodically and will be publicly shared via TALOS' Zenodo and github accounts as well as TALOS website by May 1st, 2026 (before the 2nd Revision of the project).

The DMP will be updated and adjusted regularly in line with the progress of the project. The final version of the DMP is due on November 30th, 2027 (D1.7/D7).

---

*End of the Report*