

ATRIUM – Advancing Frontier Research in the Arts and Humanities

Work Package 5
Curiosity-driven Demonstrators

Deliverable D5.1
D5.1 Interim report on Demonstrators

Funding Instrument:	Horizon Europe
Call:	HORIZON-INFRA-2023-SERV-01
Call Topic:	Research infrastructure services to support health research, accelerate the green and digital transformation, and advance frontier knowledge (2023)
Project Start:	1 Jan 2024
Project Duration:	48 months
Beneficiary in Charge:	Archaeology Data Service, University of York
Document Identifier:	10.5281/zenodo.17900562



**Funded by
the European Union**

Deliverable Information

Action Number:	101132163
Action Acronym:	ATRIUM
Action title:	Advancing FronTier Research In the Arts and hUMANities
Deliverable Number:	D5.1
Deliverable Full Title:	D5.1 Interim report on Demonstrators
Deliverable Short Title:	Interim report on Demonstrators
Document Identifier:	ATRIUM-D5.1-Interim-Report-on-Demonstrators
Beneficiary in Charge:	Archaeology Data Service, University of York
Report Version:	1.0
Contractual Date:	31/12/2025
Submission Date:	30/12/2025
Dissemination Level:	PU
Nature:	Report
Lead authors:	Émilie Pagé-Perron (ADS), Julian D. Richards (ADS)
Authors:	Kristis Alexandrou (Cyl), Georgios Artopoulos (Cyl), Anna Aslanoglou (ATHENA-RC), Anne Baillot (DARIAH), Alessia Bardi (CNR), Elton Barker (ATHENA-RC), Sheena Bassett (ARIADNE RI), Sarah Bénérière (Inria), Ceri Binding (USW), Marco Callieri (CNR), Massimiliano Carloni (OEAW), Thibault Clérice (Inria), Matej Ďurčo (OEAW/DARIAH), Tim Evans (ADS), Achille Felicetti (ARIADNE RI), Johan Fihn (SND), Maria Gavrilidou (ATHENA-RC), Guntram Geser (ARIADNE RI), Mark A. Greenwood (USFD), William Illsley (SND), Maria Ilvanidou (ATHENA-RC), Ulf Jakobsson (SND), Νίκος Καπραλός (ATHENA-RC), Joakim Larsson (SND), Olga Lečbychová (ARUB), Kateryna Lutsai (CU), Arek Margraf (PCSS), Diana G. Maynard (USFD), David Novák (ARUP), Aleksandra Nowak (PCSS), Rahaf Orabi (Cyl), Petr Pajdla (ARUB), Vayianos Pertsas (ATHENA-RC), Prokopis Prokopidis (ATHENA-RC), Pavel Straňák (CU), Maria Theodoridou (FORTH), Chara Tsoukala (ATHENA-RC), Doug

Tudhope (USW), Piotr Tylczyński (PCSS), Valentina Vassallo (Cyl), Holly Wright (ADS)
 Reviewers: Émilie Pagé-Perron (ADS), Julian D Richards (ADS), Anne Baillot (DARIAH), Megan Black (DARIAH), Toma Tasovac (DARIAH).
 Keywords: Research Infrastructure, Demonstrators, Archaeological Documentation, ATR, Information Extraction, Knowledge Graph, Digital Humanities, Image Annotation, 3D Models, Sound-Based Data, Geospatial Data, Linked Open Data (LOD), Controlled Vocabularies, Metadata Enrichment, Heritage Data, AI, NER, Multilingual NLP, Cultural Heritage, Workflow, Open Science
 Status: Final

Change Log

Date	Version	Author/Editor	Summary of Changes made
26/10/25	v0.1	Émilie Pagé-Perron	Outline
21/11/25	v0.2	All authors	First draft for most sections
6-7/12/25	v0.3	Julian D Richards	Summary, introductory and linking sections; copy-editing and harmonisation between sections; cleaning references
1-28/12/25	v0.4	Émilie Pagé-Perron, Anne Baillot, Megan Black, Julian D Richards, Toma Tasovac	Conclusion, review, proof reading, harmonisation, editing and formatting
30/12/25	v1.0	Megan Black	Final proofread & submission

Table of Contents

List of Figures	4
List of Abbreviations	6
Executive Summary	7
1. Introduction	8
1.1 Purpose and Scope of the Document	8
1.2 Structure of the Document	8
2. T5.1 Text-Based Demonstrators	9
Introduction	9
2.1 Document segmentation and text recognition of archaeological documentation (T5.1.1)	10
2.2 Information Extraction for fieldwork reports, published papers, and text from speech (T5.1.2)	15
2.3 Process knowledge graph demonstrator (T5.1.3)	28
3. T5.2 Image-Based Demonstrators	38
Introduction	38
3.1 Altamira, Mirador-based IIIF viewer	39
3.2 Interface updates at the Archaeology Data Service	40
3.3 Early mediaeval sculpture (T5.2.1)	41
3.4 Bronze Age Rock Art (T5.2.2)	42
3.5 Archival photographic collections image annotation (T5.2.3)	43
4. T5.3 3D-Based Demonstrators	47
Introduction	47
4.1 3D architectural models (T5.3.1)	48
4.2 Adoption of Historic Building Information Modelling (HBIM) (T5.3.2)	59
5. T5.4 Sound-Based Demonstrators	63
Introduction	63
5.1 Provision of sample datasets	64
5.2 The Web Application	64
5.3 Provision of final datasets	66
5.4 Next steps	66
6. T5.5 Geospatial Demonstrators	67
Introduction	67
6.1 Collaborative Map Annotation (T5.5.1)	67
6.2 Using place to connect multiple disciplines across the Arts and Humanities and beyond (T5.5.2)	71
7. ARIADNE Ontology (AO-Cat) updates	77
7.1 Introduction	77
7.2 Changes to the Ontology	77
8. Updates to the portal	82
9. Conclusion	88
References	89
Consortium	91
Disclaimer	92

List of Figures

List of Figures in Section 2: T5.1 Text-Based Demonstrators

Figure 2.1: Architecture of the OCR/HTR system with Dashboard and AP	13
Figure 2.2: OCR Service Client Dashboard	14
Figure 2.3: REST API Interface	14
Figure 2.4: Vocabulary-based subject and temporal metadata indexing pipeline	18
Figure 2.5: Sample output in HTML	18
Figure 2.6: Sample JSON output from pipeline	19
Figure 2.7: Geometric disambiguation of "Memphis" using bounding boxes	22
Figure 2.8: Czech data enrichment workflow	25
Figure 2.9: Document Page Counts Over Time by Category	25
Figure 2.10: ATRIUM Speech-based NER demo	28
Figure 2.11: Neo4j prefixed queries interface	30
Figure 2.12: Neo4j query input field	31
Figure 2.13: Neo4j KB schema	31
Figure 2.14: Neo4j Query result in table form.	32
Figure 2.15: Neo4j Query result in textual form.	32
Figure 2.16: Query result in code form.	33
Figure 2.17: Neo4j Query result displayed in graph form.	33
Figure 2.18: GraphDB Workbench interface.	34
Figure 2.19: GraphDB Query results in bar chart form.	35
Figure 2.20: GraphDB Query results in simple tabular form.	35
Figure 2.21: Query results in a pivot table.	36
Figure 2.22: Query results displayed in visual graph form	36

List of Figures in Section 3: T5.2 Image-Based Demonstrators

Figure 3.1: Altamira IIIF Viewer showing annotations on stained glass.	39
Figure 3.2: ADS interface showing the "View in IIIF viewer" button.	40
Figure 3.3: ADS query results showing "View results in Mirador" option.	41
Figure 3.4: The AMCR system pipeline for photographs	44
Figure 3.5: Typical examples of metal detecting finds from AMCR-PAS	45
Figure 3.6: An example of semi-annotated archival photograph	46

List of Figures in Section 4: T5.3 3D-Based Demonstrators

Figure 4.1: Snapshot of the EpHEMERA platform collection view.	49
---	----

Figure 4.2: Interactive implementation of the monument in PoTree.	51
Figure 4.3: Tests in 3DHOP highlighted some issues in the textures visualisation.	52
Figure 4.4: Visualisation of the HBIM reconstructed model of the Monastery	53
Figure 4.5: Interactive visualisation of the Basilica and its surroundings in 3DHOP.	55
Figure 4.6: Interactive visualization of the Kampanopetra Basilica in 3DHOP.	56
Figure 4.7: Snapshots of the virtual reconstruction visualisation.	56
Figure 4.8: 3DHOP visualization with a “see-through” mode	57
Figure 4.9: Metadata alignment for the aggregation into the ARIADNE portal	58
Figure 4.10: General-purpose information benchmark	60
Figure 4.11: Pilot Building Case	61
Figure 4.12: In Urban Periscope Portal	61
Figure 4.13: In the ADIADNE PORTAL	62

List of Figures in Section 5: T5.4 Sound-Based Demonstrators

Figure 5.1: Voice recording web app interface prompt.	64
Figure 5.2: User dashboard for Context Forms.	65
Figure 5.3: Interface for adding a new context sheet using voice commands	65

List of Figures in Section 6: T5.5 Geospatial Demonstrators

Figure 6.1: Annotation of the Charta of Greece in Recogito Studio.	68
Figure 6.2: Resource distribution in the ARIADNE Portal.	72
Figure 6.3: Östergötland pilot data plotted in the GIS interface.	74
Figure 6.4: Excavated gravestone from St James's Burial Ground (Context 101035).	75

List of Figures in Section 7: ARIADNE Ontology (AO-Cat) changes

Figure 7.1: Model for images copied into the ARIADNE Cloud.	80
Figure 7.2: Model for images referred to in the ARIADNE Cloud.	81

List of Figures in Section 8: ARIADNE Portal changes and upgrades

Figure 8.1: ARIADNE Portal homepage	82
Figure 8.2: Screenshot of the online Portal User Manual	83
Figure 8.3: New sorting options in the search results	84
Figure 8.4: The ARIADNE map viewer display	84
Figure 8.5: New Data Type filters in the search interface	85
Figure 8.6: Filer showing multiple options selected	86
Figure 8.7: IIIF Manifest and Image display in the Portal	87

List of Abbreviations

AAT	Getty Art and Architecture Thesaurus
ADS	Archaeology Data Service, University of York (UoY-ADS)
AMCR	Archaeological Map of the Czech Republic
AO-Cat	ARIADNE Ontology Catalogue
API	Application Programming Interface
ARUB	Institute of Archaeology of the Czech Academy of Sciences, Brno
ARUP	Institute of Archaeology of the Czech Academy of Sciences, Prague
ASR	Automatic Speech Recognition
ATR	Automatic Text Recognition
ATRIUM	Advancing Frontier Research in the Arts and Humanities
BIM	Building Information Modelling
CAA	Computer Applications and Quantitative Methods in Archaeology
CH	Cultural Heritage
CNR	Consiglio Nazionale delle Ricerche
CU	Charles University
CYI	Cyprus Institute
FBC	Federacja Bibliotek Cyfrowych (Polish Digital Libraries Federation)
HBIM	Historic Building Information Modelling
HTR	Handwritten Text Recognition
IFC	Industry Foundation Classes
IIIF	International Image Interoperability Framework
JSON	JavaScript Object Notation (file format)
KB	Knowledge Base
LOD	Linked Open Data
LLM	Large Language Model
LNEC	Laboratório Nacional de Engenharia Civil
NER	Named Entity Recognition
NLP	Natural Language Processing
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCR	Optical Character Recognition
OEAW	Austrian Academy of Sciences
PAS	Portable Antiquities Scheme
PCSS	Poznań Supercomputing and Networking Center
SSHOMP	Social Science and Humanities Open Marketplace
SND	Swedish National Data Service
SKOS	Simple Knowledge Organization System
TEI	Text Encoding Initiative
USW	University of South Wales
UFSD	University of Sheffield
WADM	Web Annotation Data Model
WP	Work Package

Executive Summary

ATRIUM brings together 30 partners from 12 countries, including four major European research infrastructures – DARIAH, ARIADNE, CLARIN, and OPERAS – to strengthen the foundations of arts and humanities research across Europe.

This report provides an interim review of progress with the demonstrators developed in Work Package (WP)5 ‘Curiosity-driven Demonstrators’ at the half-way point in the ATRIUM project. The demonstrators are based on real world arts and humanities research questions, and they build on the workflows that are being developed in WP4 ‘Providing Enhanced Workflows for Frontier Research in the Humanities’, as well as taking account of the new audiences explored in WP2 ‘Communication, Dissemination and Impact Evaluation’. The workflows and demonstrators were selected to cover five primary data types used in the arts and humanities: text, images, 3D, sound, and geospatial data. The demonstrators focus on the discipline of archaeology given the exceptional variety of data types it employs, although they are of wider applicability. Test data for the workflows has been provided by the archaeological partners. Half-way through the project the majority of the workflows are at an advanced stage of development, although some require further testing. The nature of each of the demonstrators, and the specific case studies, have all been agreed, and several of the demonstrators are in a prototype stage.

The archaeological metadata aggregator provided by the ARIADNE Portal is the platform for the majority of the demonstrators, and a number of enhancements have been made to both the platform and the underlying ontology, improving usability, and facilitating integration of the demonstrators. In the second half of the project the demonstrators will be completed, enhanced metadata will be uploaded to the ARIADNE Knowledge Base, and links will be embedded in the ARIADNE portal.

1. Introduction

1.1 Purpose and Scope of the Document

This report provides a mid-term review of the curiosity-driven demonstrators developed in WP5 at the halfway point of the ATRIUM project. It documents the progress based on the Description of Work, assesses the current maturity of each demonstrator, and identifies risks and planned next steps to support any remaining decisions to be made for the project's second half.

We distinguish between workflows (WP4) and demonstrators (WP5), though we also acknowledge that some significant natural overlap exists. Workflows include both methodological sequences and software pipelines designed to be generic, portable, and applicable across datasets and contexts. Demonstrators are targeted, case-specific implementations that apply these workflows to particular research questions and corresponding selected datasets. During the implementation of the demonstrators, we fine-tune the pipelines applied for our case studies while maintaining the core reusability for the wider community. This report often describes parts of the WP4 workflows to support the clarification of the demonstrators work. The workflows and demonstrators were selected to encompass the variety of data types used in arts and humanities research. We focused on archaeology given its exceptional data diversity, though the approaches have wider applicability. Most demonstrators are showcased through the ARIADNE RI online portal for archaeology. [The test datasets](#) were supplied by archaeological partners in ARIADNE, while technical development was supported by partners from DARIAH and CLARIN. The demonstrators also take into account the new audiences explored in WP2.

1.2 Structure of the Document

This document is structured according to the five broad data types selected for the workflows and demonstrators. These embrace the primary data types used in archaeological recording: text, images, 3D, and geospatial. To these we added sound as a novel means of site recording in which there is experimentation. [Section 2](#) describes progress on task 5.1, the text-based demonstrators; [Section 3](#) covers T5.2, the image-based demonstrators; [Section 4](#) is concerned with T5.3, the 3D demonstrators; whilst [Section 5](#) reviews progress on the sound-based demonstrator. Finally, [Section 6](#) reviews progress on the geospatial demonstrators. [Section 7](#) then outlines the changes made to the ARIADNE RI ontology, AO-Cat, which was necessary to accommodate the demonstrators within the portal. [Section 8](#) itemises the enhancements made to the portal interface itself. The deliverable is concluded in [Section 9](#).

2. T5.1 Text-Based Demonstrators

Introduction

Archaeologists create text documents in multiple contexts: when they record their site observations in the field; during post-excavation as they write-up their results (often in unpublished 'grey literature' reports) and during publication in learned journals and monographs. Repositories routinely curate examples of all these varieties of documents, but frequently lack comprehensive metadata to aid resource discovery and research re-use. The text-based workflows address this challenge by enabling segmentation, recognition, and semantic enrichment of heterogeneous textual sources.

Halfway through the project, the WP4 workflows are in an advanced state. The text-based demonstrators apply these workflows to datasets held by national repositories, notably the [Archaeology Data Service, University of York \(ADS\)](#) and the [Czech national repositories at Archaeological Map of the Czech Republic \(AMCR\)](#). As part of the preparatory work for implementation of the demonstrators, the team has successfully created [Open Archives Initiative Protocol for Metadata Harvesting \(OAI-PMH\)](#) harvesters from ARIADNE to ADS and AMCR so that this metadata can be automatically updated as new resources are added.

This chapter presents three distinct text-based demonstrators: (1) text recognition of archaeological documentation ([section 2.1](#)), which focuses on segmentation and Automatic Text Recognition (ATR) workflows to convert heterogeneous document collections into machine-readable text; (2) information extraction for fieldwork reports and published papers ([section 2.2](#)), which applies Natural Language Processing (NLP) to extract subject, temporal, and spatial metadata from grey literature, journal articles, and text from speech, in English and Czech; and (3) process knowledge graph demonstrator ([section 2.3](#)), which creates and visualises knowledge graphs from research processes documented in archaeological publications.

During the next 20 months the workflows and their application will be further refined and tested, with the next steps focusing on ingesting the enhanced metadata prepared to the ADS and AIS-CR repositories, and then aggregating it to the ARIADNE portal for improved findability.


2.1 Document segmentation and text recognition of archaeological documentation (T5.1.1)

2.1.1 Overview

This task applies page segmentation and Automatic Text Recognition (ATR) workflows to convert our chosen large and heterogeneous archaeological document collections into machine-readable text and structured layouts as a basis for later enrichment and reuse. As part of the demonstrators, over 115,000 PDF files from ARUP & ARUB were processed using OCR to generate ALTO-XML files. Improved segment annotation guidelines were prepared, and preliminary segmentation tests have been performed on sample datasets from Czech and UK institutions.

Workflows

Segmentation and ATR: <https://marketplace.sshopencloud.eu/workflow/sS4gSB>

Dashboard and API for ATR:  T4.1.1 Automatic Text Recognition Workflow DRAFT

2.1.2 Provision of datasets

With regards to Czech data, ARUP has provided 66,991 PDF files containing textual documents, generally scanned from analogue, dated from 1920 to 2022 and comprising 649,508 individual pages. The Institute of Archaeology of the Czech Academy of Sciences, Brno (ARUB) provided 48,167 PDF files comprising 641,082 individual pages.

As part of the workflow, and as a basis for the demonstrator, these documents were processed into ALTO-XML files (one per PDF) containing recognised page elements and textual data. These ALTO files were supplemented with corresponding TXT versions and, in the case of ARUP, also PDF files with a textual layer. For now, the data is stored locally and used for testing and improving the workflow. But in 2026, we will consider publishing the ALTO files as an alternative distribution to the original PDF files which are all already stored and published in the [AMCR repository](#).

ARUP and ARUB provided a purposefully [selected set of archetypal pages](#) that are typical of the documents held by these institutions. These were to be used to test the segmentation workflow and establish an appropriate methodology to cover all the important elements present on the pages, such as forms, figures, tables, stamps and text. ADS also provided samples composed of a range of archaeological context sheets and accompanying metadata. They were pre-processed to [single pages](#) based on the processing requirements. The final dataset from ADS will comprise a similar sample but will be larger in size. All the files are already archived at ADS and available publicly for download under a CC-BY licence.

For this demonstrator, the final dataset for the Czech data will be supplemented with more recent files held in the AMCR (documents deposited with AMCR between now and the end of the ATRIUM), and the underlying pipeline of the workflow will be operationalised to automatically process new textual files deposited at the provider using local or distributed services accessible via APIs (<https://api.aiscr.cz/>). This means that the workflow will be constantly reused automatically at AMCR after the end of ATRIUM.

2.1.3 Segmentation

The ATR workflow uses object detection (Clérice et al. 2025); it relies on two separate tools: the computer vision platform [Roboflow](#) and the ATR platform [eScriptorium](#), itself relying on the [kraken](#) ATR engine.

For testing the workflow and starting the work on the demonstrator, representative documents were provided, as described above. These documents illustrate the core challenges their providers face: dealing with the structural complexity of archaeological forms, accurately detecting images, figures and tables, and handling documents which combine printed and/or typewritten text with handwritten text. As such, the workflow's goal is to process these documents into machine-readable content with a fine-grained typology of content regions in plain text and ALTO-XML formats for use in the demonstrator.

The primary layout analysis task of the demonstrator is carried out using the object detection model derived from the Layout Analysis Dataset with SegmOnto (LADaS) (Clérice et al. 2024). The [LADaS dataset](#) is composed of over 8,000 images of various types of documents (e.g. research articles and theses, novels, plays, poetry, dictionaries, sales catalogues) from the 17th to the 21st century, illustrating a great variety of layout across time. The dataset was annotated using 44 classes, described in the [LADaS Annotation Guidelines](#) (Janès et al. 2025) to allow for a fine-grained semantic interpretation of content.

In part due to the challenges raised by the ATRIUM data, the team revised and improved the LADaS Guidelines in order to incorporate a typology for documents containing forms, and reduce, or even eliminate, ambiguous cases in layout annotation. This redesign required both internal collaboration at Inria with LADaS 1.0's originating project ([DEFICOLaF](#)), and external consultation with partners (including, but not limited to ADS, ARUB and ARUP). The updated LADaS 2.0 Guidelines are currently being validated through the complete review of the associated 8,000+ images dataset.

Following this update, in 2026, we will strategically reannotate a sufficient amount of data to allow the existing model to perform accurate pre-annotation on archaeological documents. This preparation is intended to optimise the efficiency of the subsequent joint annotation campaign with the data providers, as well as limit friction with an otherwise large set of annotation types.

Due to internal administrative issues at Inria, the planned recruitment of a computer vision specialist was delayed by 6 months from the initial project timeline. Dr Benjamin Kiessling, the primary developer and curator of the kraken ATR engine, was integrated into the Inria team. In the interests of promoting open source tools, the decision was made to embed an object detection model directly into kraken, for which a functional [prototype](#) is in progress. However, given the reliance on eScriptorium by the community, and more widely on kraken, the prototype requires extensive testing before its release as a stable version to not disrupt research workflows that are currently in use. Integrating the object detection approach will simplify the original workflow as it will allow all annotation work to be conducted within eScriptorium only, instead of two different platforms (Roboflow and eScriptorium). This will also remove any technical requirements beyond the use of a graphical user interface (GUI) because eScriptorium will then host both the training of and prediction using object detection models within its interface.

Inria is actively working on improving eScriptorium and its segmentation GUI, in alignment with ATRIUM's commitment to encourage the use and development of open source tools. However, unforeseen development setbacks have caused a delay in the release of the newest version of eScriptorium, consequently impacting the overall demonstrator preparation timeline.

Next steps

Workflow:

- Release of eScriptorium's new User Interface (Q1 2026)
- Stable kraken 7.0 with Object Detection (Q1 2026)
- Integration of kraken 7.0 into eScriptorium (Q1/Q2 2026)

Demonstrator:

- Annotation workshop and campaign with the data providers (Q2 2026)
- Monitoring of the demonstrator evolution (2026–2027)

2.1.4 ATR Dashboard and API Demonstrator

Summary

To support batch processing for OCR of documents, PCSS developed a dual solution combining a user-friendly dashboard for individual tasks and a REST API for automated mass processing. PCSS is modifying an existing dashboard and has implemented a first version of the API for this workflow, to support the associated demonstrator.

This architecture is expected to integrate Tesseract and Kraken engines through a common Java-based interface, enabling both OCR and Handwritten Text Recognition (HTR) functions with future scalability and page layout recognition support. This workflow is an essential underlying component for the realisation of the text-based demonstrators.

Workflow

After conducting a needs analysis, the PCSS dashboard was adapted for the workflow and its associated demonstrator since users will need to process individual documents. Mass processing will also be required so a REST API is under development, where the dashboard can still be used to monitor the progress of batch processing.

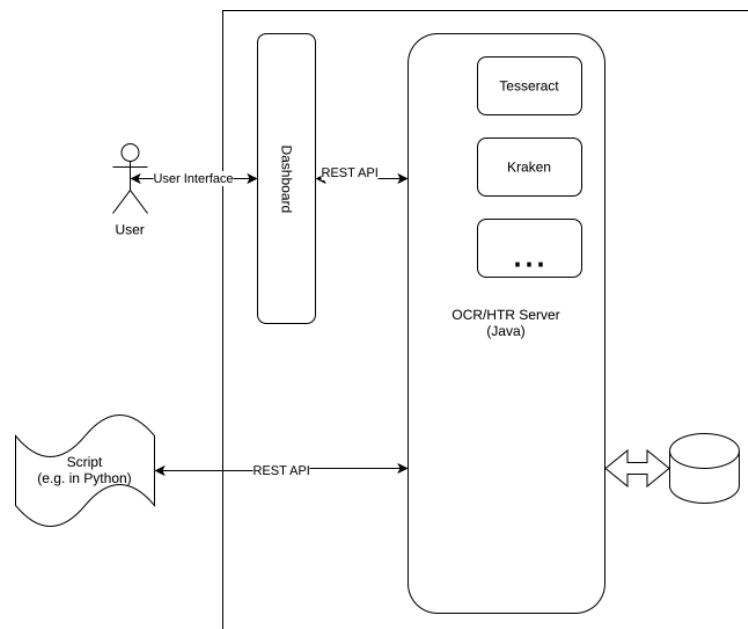


Figure 2.1: Architecture of the OCR/HTR system with Dashboard and API.

Both the dashboard and the API are made flexible enough so they can use different engines. First to be implemented are tesseract and kraken, the two engines covering the most required features of high-quality OCR and HTR. More engines will be included if there is a real need, and if time allows. Page layout recognition was another challenge to address because many archaeological documents are forms, partly printed and partly handwritten. Kraken is used for this task, OCR is then applied to individual segments. Page layout recognition will be built into the OCR mechanism as a preliminary stage of document processing. In this case, HTR segmentation is not

necessary, as Kraken does it automatically. After adopting the above assumptions, the system will have the architecture shown in Figure 2.3.

So far, we have developed a preliminary version of the dashboard, and a server has been written to offer the REST API. Both components were written in Java. [The REST API](#) (Figure 2.3) has been designed and [sample programs in Python](#) that show how to use it were developed. The description of the API workflow can be found here:

[T4.1.1 Automatic Text Recognition Workflow DRAFT](#)

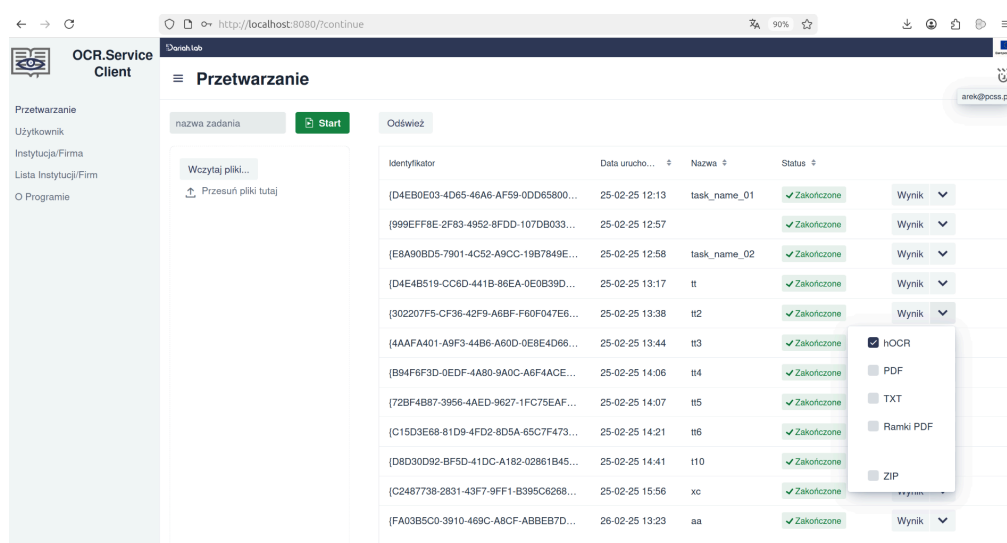


Figure 2.2: OCR Service Client Dashboard

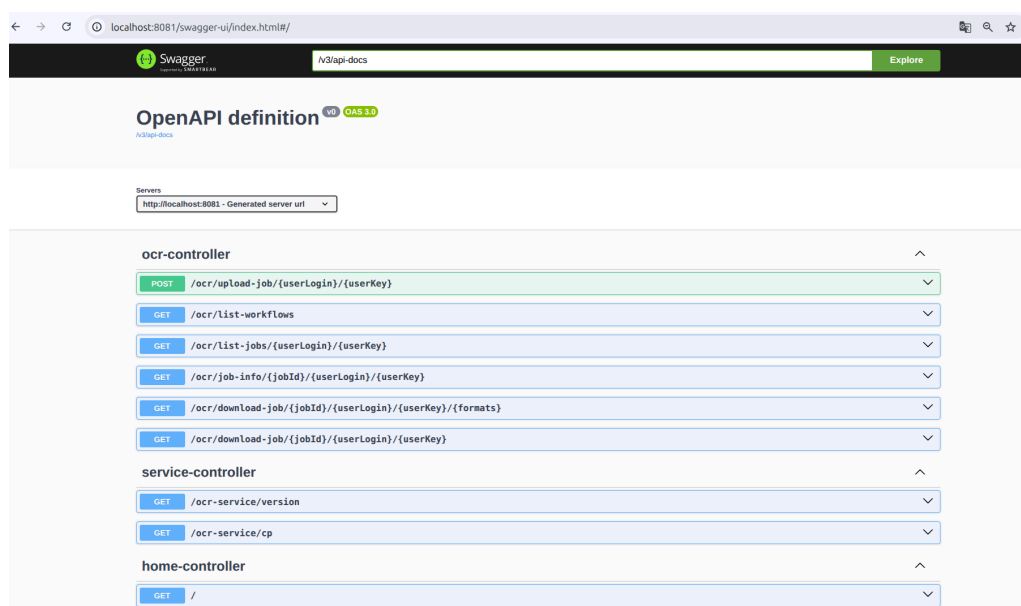


Figure 2.3: REST API Interface

Next steps

Integration of Tesseract and Kraken is currently ongoing. After integration, users of the dashboard and API will be able to process jpeg, png, and tiff files to obtain segmented and ATRed ALTO-XML results, with accompanying text files.

Further work will focus on ensuring full scalability of the solution through parallelisation and security through appropriate user authorisation/ authentication. The dashboard interface will be rendered more accessible and will better display monitoring of ongoing tasks.

2.2 Information Extraction for fieldwork reports, published papers, and text from speech (T5.1.2)

2.2.1 Overview

The T5.1.2 Information Extraction Demonstrator addresses the challenge of effective information retrieval from archives of unpublished fieldwork reports (or grey literature), together with libraries of digitised back-runs of local history and archaeology journals. Existing subject metadata for these resources is sparse and inconsistent and indexing the metadata manually is an extensively labour-intensive process. The demonstrators apply and operationalise the workflows developed in T4.1.2 using the English and Czech languages, as examples of one widely used language and one less commonly implemented.

Work to date on the T5.1.2 Demonstrator has applied the subject, temporal, and spatial information extraction workflows to ADS Library grey literature and journal samples, generated enriched metadata for “What-When-Where” facets, and iteratively tuned ranking and configuration using ADS expert feedback. In parallel, the Czech and speech-based Demonstrators have operationalised their respective workflows on large AMCR and context-sheet samples by classifying pages of archival documents, and standing up a tested speech-based NER demonstrator with an expert correction interface for ASR-derived context sheet descriptions.

2.2.2 The English Language Demonstrator

The input data resources for the English language Demonstrator are derived from the [ADS Library](#), which holds both bibliographic metadata records and Open Access (full text) copies of published and unpublished documents relating to archaeology and heritage. The ADS Library is regularly updated with new collections from publishers and

fieldwork reports deposited via the [OASIS](#) system for reporting investigations into the historic environment. It also holds scanned documents and oral material.

The Demonstrator draws on various information extraction software components that can be combined via a common JavaScript Object Notation (JSON) output format to automatically generate metadata elements that describe: What, When and Where (subject data, spatial data, and temporal data). These components are based on the workflows developed in T4.1.2 as explained in the below sections [Subject and Temporal metadata enrichment](#) and [spatial metadata enrichment](#).

Subject and Temporal Metadata Enrichment

The [workflow for subject and temporal metadata elements](#) has significantly extended and refined existing tools for vocabulary-based metadata enrichment developed in the ARIADNE and ARIADNE Plus projects and other work (Binding and Tudhope 2023; 2024a). These tools are implemented as a multilingual pipeline based on the spaCy open-source Natural Language Processing (NLP) library, augmented with rule-based patterns expressed as Python modules and various specialised pipeline components.

For T5.1.2 Demonstrator purposes, the pipeline subject component employs the [FISH Archaeological Object Thesaurus](#) and the [FISH Thesaurus of Monument Types](#) but other vocabularies are also possible, typically available [as Simple Knowledge Organization System \(SKOS\) Linked Open Data](#). For example, the Getty Art and Architecture Thesaurus (AAT) is directly available in the pipeline (the Monuments and Objects thesauri were mapped to the AAT as part of the ARIADNE projects).

For temporal metadata, archaeological named periods are taken from the Historic England '[Archaeological and Cultural Periods](#)' [PeriodO](#) authority, "Historic England", but it is also possible to use other PeriodO authorities. In addition to named periods (e.g. Early Mesolithic), textual numeric patterns for various categories of temporal expression are also extracted by the pipeline such as ordinal named or numbered centuries, year spans, and century spans, etc. Vocabulary lookup matches are augmented by pre-processing specialised components such as token normalisation and various NLP operations including tokenisation, Part-Of-Speech tagging, and lemmatisation.

ATRIUM refinements have included the extension of the thesaurus entry vocabulary for automatic indexing (e.g. alternate terms), normalisation of whitespace and punctuation elements, and negation detection. Additionally, a significant extension has seen the detection of pairings of temporal and subject elements (e.g. medieval brooch). The pipeline underlying the workflow is implemented via Python Notebooks and output

metadata is produced in various formats (text, JSON, HTML), including the recommended metadata and sets of vocabulary concepts (with URI identifiers where available) together with the start/end character positions for the associated text spans. Bulk processing scripts are available, together with [Python notebooks](#) that illustrate usage and highlight significant aspects of the functionality as well as the open source [pipeline](#) (Binding, n.d.). All of these refinements increase the quality of the additional metadata produced as part of the demonstrator.

Sets of [sample data](#) for the current stage of the Demonstrator from both OASIS (grey literature) reports and heritage journal back runs were made available by ADS. The OASIS reports included existing metadata with abstracts and it was decided after analysis that the abstracts were the appropriate focus for OASIS metadata enrichment, as they emphasise the key findings by the report authors. Some of the more recent journal articles come with substantial abstracts but many of the older articles do not; articles tend not to follow any standard structuring style. Hence it was judged appropriate to focus on the full text for the journal articles. The [ADS sample](#) of full text PDF articles is available, together with a conversion to corresponding text documents by task partners USFD. Processing the full text presents challenges as the pipeline will generate large numbers of subject and temporal annotations. This full set of automatic metadata may still address use cases concerning highly specific search statements. However, for more general full text cross search use cases in the Demonstrator, current work is refining a mechanism for ranking and prioritising the vocabulary matches to support controlled vocabulary subject indexing and information extraction for the ADS Library and the ARIADNE Portal Demonstrator. A variety of methods are combined, including concept frequency, position within the document structure, and contextual indicators of authorial emphasis and significance, drawing on feedback from ADS domain experts. A ranking formula is based upon a weighted average of the different factors. Weightings for the different factors are adjustable parameters at pipeline configuration for a particular input collection. For example, analysis shows that the title is likely to be highly informative for journal articles but less so for OASIS reports. Similarly, contextual indicators of significance are particularly important for full text journal articles and somewhat less so for OASIS Abstracts. Figure 2.4 shows the general pipeline.

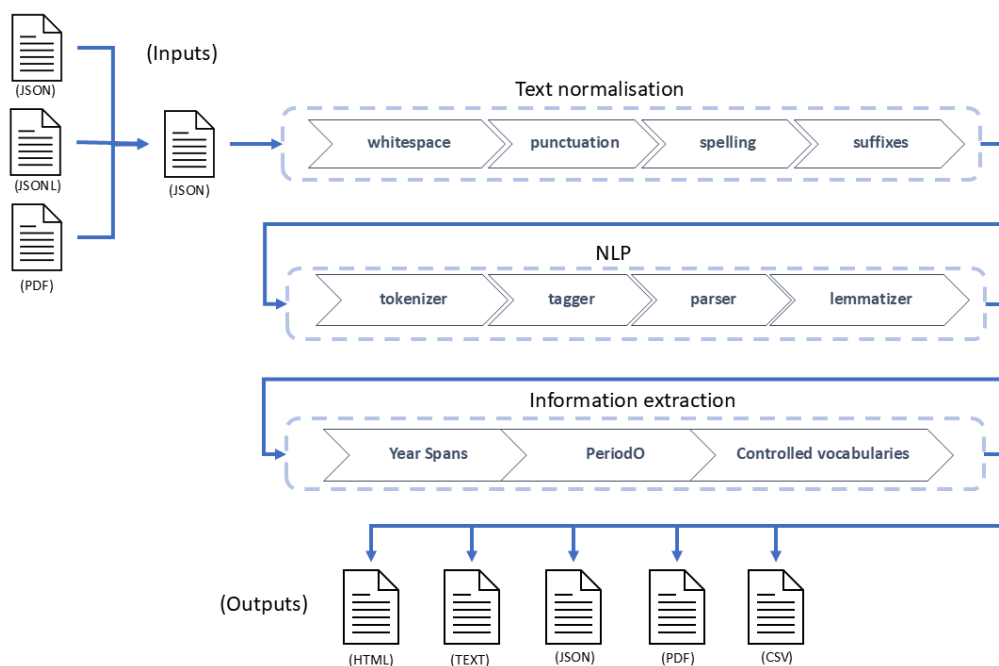


Figure 2.4: Vocabulary based subject and temporal metadata indexing pipeline

The pipeline has been developed iteratively and refined via formative evaluation. Sample output from the pipeline is shown in Figure 2.5 with colour coded entity types and Figure 2.6. Figure 2.5 also shows detection of both named periods and numeric year spans, including radiocarbon dating. Text output formats can include metadata about the particular pipeline process and date, concept pairings, negations, and diagnostic information, etc.

activity spanning from the **Early Neolithic** through to the **Medieval** period . The **Early Neolithic** presence was limited to the heavily

truncated remnants of a post - built **structure** and a **waste - pit** radiocarbon dated to **3947 - 3715 BC** . Succeeding activity spanned from

the **Early Beaker** period to the **Early Bronze Age** and included a **Beaker ' midden ' pit** dated to **2398 - 2146 cal BC** , and then a funerary

ring - ditch monument which enclosed four **inhumations** and two **cremations** , and had been developed over five distinct phases

radiocarbon dated from **2026 - 1896 BC** through to **1869 - 1621 BC** . A contemporary or subsequent phase of what appears to be domestic

occupation immediately adjacent to the **ring ditch** took place as represented by a small rectangular **building** dated to **2133 - 1938 cal BC**

with possible **midden pits** , and also a **roundhouse** , the latter either at the same time or possibly at some stage in the **Early Bronze Age** .

Iron Age activity , sub - divided into three broad phases , demonstrated a change in land - use , as attested by the presence of a

Figure 2.5: Sample output in HTML

```
{
  "text": "The ditch itself cannot be dated in this slot, but the upper deposits (202006) (202008),
  which represent the disuse and deliberate backfilling of the feature can be dated to the Roman period
  - this also follows with the disuse layers seen in the continuation of Grim's Ditch currently being
  excavated in the \"yellow brick road\" area; although here the alignment of the ditch is altered.",
  "spans": [
    {
      "start": 4,
      "end": 9,
      "token_start": 1,
      "token_end": 1,
      "label": "FISH_MONUMENT",
      "text": "ditch",
      "id": "http://purl.org/heritagedata/schemes/eh_tmt2/concepts/70351"
    },
    ...,
    {
      "start": 178,
      "end": 190,
      "token_start": 38,
      "token_end": 39,
      "label": "PERIOD",
      "text": "Roman period",
      "id": "http://n2t.net/ark:/99152/p0kh9ds9nsj",
    },
    ...
  ],
}
```

Figure 2.6: Sample JSON output from pipeline

A pilot exercise performed before the start of ATRIUM has been published (Binding and Tudhope 2024) as well as a [report](#) on a summative, intellectual evaluation of the pipeline as of November 2024. This evaluation focused on results from the ADS sample of OASIS report abstracts, considering the correctness of each vocabulary annotation (and pairings) and whether any vocabulary concepts were missed. Evaluation took into account whether the appropriate sense of vocabulary concepts was identified. Details are discussed in the report, including the vast majority of vocabulary annotations (and pairings) corrected with periods. The incorrect annotations were mostly due to homonyms usually associated with a small set of terms that could be placed in a stopword glossary and a small set of disambiguation issues, arising from an unsuccessful attempt to deal with parenthetical qualifiers (subsequently reversed). The small number of missing annotations were due to limitations in the entry vocabulary and to incorrect tagging by the spaCy POS pipeline component. Evaluation findings, including extension of the entry vocabulary and domain tailoring of the vocabularies, have been incorporated into subsequent iterations of the pipeline. The ATRIUM extensions to the entry vocabulary have also been communicated to the FISH Terminology Group for consideration in their future releases.

The work has been presented at the Computer Applications and Quantitative Methods in Archaeology (CAA) conference in 2025, a key event for digital archaeology. Both the subject (Binding and Tudhope 2025a) and temporal (Binding and Tudhope 2025b)

components were reported on separately. The pipeline was presented at a workshop at ADS on 27th of November 2025 with an initial ranking configuration and this will be refined in the light of expert user feedback.

Next steps

Workflow:

Plans for future work on the workflow include refining the mechanism for ranking and prioritising vocabulary-based information extraction to support controlled vocabulary subject indexing, and developing configuration parameters that allow users to tailor the pipeline to particular datasets. The Python Notebooks documenting the pipeline will be further extended and the associated workflow will be published on the Social Sciences and Humanities Open Marketplace, together with a batch operation mode to generate subject and temporal metadata for datasets of OASIS reports and ADS Library journal backruns.

Demonstrator:

For the demonstrator, ADS will review the batch-generated output with the development team and investigate appropriate selection strategies (cutoffs) for the ranked subject metadata and paired compound concepts, before adding the resulting metadata to the corresponding publication records. This enriched metadata will then be ingested into the ARIADNE Portal via regular harvesting of the ADS OAI-PMH API, with FISH Monuments and Objects concepts mapped to the Getty AAT.

Named Entities Recognition of Place

The Demonstrator also enriches metadata for ADS grey literature and journals through the workflow that extracts named entities related to place, specifically locations or geographical features which define the 'Where' facet, and their subsequent disambiguation against relevant Linked Open Data (LOD) resources.

Compliance with Original Plan and Initial Challenges

The original plan outlined the necessity of a robust methodology for NER and entity linking. Initial work focused on adapting the existing geoparser [Mordecai 3](#) to meet the specific requirements of the archaeological domain. While GeoNames is a comprehensive resource, it was determined to be unsuitable in isolation due to its low coverage of ancient and historic place names. The initial intention to substitute the GeoNames index with data from more appropriate LOD resources (e.g. Pleiades, World Historical Gazetteer) encountered two primary technical barriers:

- **Model Dependency:** The machine learning disambiguation component was heavily reliant on specific, proprietary fields within the GeoNames schema, making adaptation for other datasets technically complex.
- **Extraction Limitation:** The use of a generic spaCy model for the initial extraction phase proved inadequate. It failed to achieve requisite recall, particularly when dealing with ancient place names that lie outside the scope of its general training data.

Gazetteer-Based Extraction and Geometric Disambiguation

To mitigate these limitations and fulfill the ultimate goal of high-coverage entity linking, the process was fundamentally re-engineered. The revised methodology reverses the direction of dependency, utilising the LOD datasets as the primary source for entity recognition, rather than solely for post-extraction disambiguation.

Extraction Phase: LOD-to-Gazetteer Conversion

The team leveraged the completeness of relevant LOD resources (e.g. Pleiades) to ensure high-recall recognition of target entities. A [standard GATE gazetteer](#) was constructed directly from the LOD. For each entity (e.g. an ancient place), all associated names (e.g. attested and romanised fields from Pleiades) are extracted. These names are indexed in the gazetteer, associating the resulting annotation directly with the entity's unique identifier, time period, and geographical coordinates. This ensures that every name present in the foundational LOD can be successfully identified in the input text, resolving the recall issue faced with the generic spaCy model.

Linking Phase: Geometric Disambiguation

The gazetteer-based extraction inevitably results in ambiguity (e.g. "Memphis" could refer to sites in Tennessee or Egypt). To resolve this, a geometric approach was employed, grounded in the "one sense per discourse" principle often used in Word Sense Disambiguation:

- **Hypothesis:** The text being processed (a sentence, paragraph, or report section) discusses entities that are geographically proximate.
- **Mechanism:** Given a set of ambiguous candidates extracted from a defined text segment, the system selects the specific configuration of candidates that minimises the size of the axis-aligned bounding box (AABB) containing all their associated geographical coordinates.



Figure 2.7 : Example sentence, where the two geographic locations (highlighted in red) are successfully resolved by the geometric disambiguation method

In the example above, the geometric approach successfully resolves "Memphis" to the ancient Egyptian capital due to its geographical proximity to both Alexandria and Thebes, resulting in the smallest bounding box.

The current working iteration of the new methodology of this workflow is available for testing and demonstration via the [GATE Cloud platform](#) and is also available in the following repository: <https://github.com/GateNLP/atrium-geoparsing>.

Presentation at the Lorentz Workshop

The work was presented at the [Lorentz Workshop: Enriching Digital Heritage with LLMs and Linked Open Data](#), held at the Lorentz Center in Leiden, Netherlands. This venue provided a targeted audience of international experts in digital humanities, computational linguistics, and LOD infrastructures, ensuring the approach was scrutinised by relevant peers. The [presentation](#) detailed the shift in strategy from conventional NER (Mordecai 3 adaptation) to the novel high-recall, LOD-driven gazetteer construction and subsequent geometric disambiguation method. This dissemination activity fulfilled a critical project objective by validating the technical approach and establishing visibility, connecting the entity linking tasks with relevant discussions concerning Large Language Models (LLMs) and their future utility within the digital heritage domain.

Future Work and Strategic Directions for the workflow

Building on the foundation of the geometric disambiguation approach, the subsequent work will focus on enhancing system robustness, expanding data coverage, and introducing temporal constraints to elevate linking accuracy. The next steps for this task are formally defined as follows:

- Refinement of Gazetteer Construction: The methodology for building the GATE gazetteer from the underlying LOD sources will be significantly refined. This involves developing specific data cleansing and harmonisation routines to mitigate the inherent "noise" and structural incompatibilities arising from using LODs not originally designed for direct gazetteer input.
- Expansion of LOD Integration: The scope of entity coverage will be broadened through the strategic inclusion of additional Linked Open Data resources. This expansion will target specialised datasets relevant to specific periods or geographical regions in archaeology, thereby enhancing the system's overall recall and utility.
- Incorporation of Temporal Disambiguation: The current geometrical disambiguation model will be augmented with a temporal dimension. This will leverage the time period information associated with the entities in the LOD to introduce temporal constraints into the disambiguation process, further refining the selection of the correct entity when multiple possibilities exist.
- Development of a Validation Corpus: To enable robust evaluation and demonstration, a dedicated Gold Standard dataset will be collated. This corpus of annotated archaeological reports and "grey literature" will serve as the essential benchmark for quantitative testing of the approach's performance, ensuring transparency and measurability of the final system.
- Test other historical place name gazetteers, such as [English Historical Place-Names](#).

2.2.3 Textual data enrichment - Czech data

Summary

The Czech demonstrator uses content-aware page classification, selective text processing, and vocabulary-driven NLP to improve metadata quality for large collections of heterogeneous, historical archaeological documents, enabling more accurate indexing and discovery through the AMCR repository and the ARIADNE Portal.

Published workflow

<https://marketplace.sshopencloud.eu/workflow/0xSpVP>

The **workflow** applied in the Czech demonstrator combines content-specific page classification, selective processing of textual data and vocabulary-driven natural language processing (NLP) to improve the quality of metadata in a Czech digital repository of archaeological fieldwork, AMCR, and related discovery services. All

enriched data will also be indexed in the ARIADNE Portal to make it easily accessible to the international research and heritage community.

For the Czech **demonstrator**, the preliminary outputs described in T4.1.1 were used as the main input. These were acquired in full in March 2025 for ARUP and in November 2025 for ARUB. Meanwhile, the team worked directly with the original PDF files, experimenting with visual-based classification of page contents.

The collections held by ARUP and ARUB consist of scans of paper-based materials with inconsistent or limited metadata. This legacy material, including handwritten notes, machine-typed reports, photocards, maps, drawings and mixed-content pages, presents two problems: firstly, it is valuable but difficult to find; and secondly, modern off-the-shelf recognition tools often fail on century-old or repeatedly rescanned pages exhibiting folds, stains, noise and atypical layouts. To address these issues, the team has developed workflows that first categorise page images according to their main content (e.g. tables, photographs, drawings/maps, plain text, mixed/handwritten cut-outs). Knowing a page's content enables the selective application of specialised tools, saving computing power and improving the quality of the results.

The team compiled and labelled a dataset comprising almost 50,000 scanned pages, categorised into 11 semantic content classes. To capture both layout and visual cues for image-based page sorting, we fine-tuned a range of backbones, including transformers (DiT, ViT), convolutional networks (EfficientNet2, RegNetY) and hybrid image-text models (CLIP). Model evaluation included accuracy figures and confusion matrices for each class (predicted versus true), and the trained checkpoints and evaluation scripts were published to enable reproducibility. The accompanying training code, README and prediction utilities are publicly available (prepackaged for ease of use) at: <https://github.com/ufal/atrium-page-classification>.

The five best-performing models were used to process all ARUP data (649,508 pages) by Charles University (CU) and classify it with this tool, resulting in a CSV file containing five class predictions per page. Working with this combined information enables the team to navigate the archival datasets more effectively, as demonstrated by the chart in Figure 2.8.



Content-specific classifiers were found to provide a practical systemisation of collections, substantially reducing unnecessary processing. Pages routed to targeted tools require fewer retries and produce higher-quality outputs than a one-size-fits-all pipeline. We also found that contemporary Document Layout Analysis (DLA) frameworks such as DeepDoctection, which uses Google's Tesseract for OCR and Facebook AI Research's detectron2 for structured data recognition, often misidentify tables and miss text blocks on heavily degraded archival pages when trained on clean digital PDFs. This confirms the value of a front-end content classifier tuned to historical scans to gate further processing.

The next step was to measure language and quality characteristics using the generic T5.1.1 results (see above), i.e. the ALTO XML outputs, with LanguageID and a causal-LM perplexity probe. This was done to reliably filter extracted text from noisy results using distilgpt2. Noisy handwritten segments were flagged for deselection from further processing, or for processing via more specialised HTR (e.g. finetuned Kraken/PERO-style workflows) or manual transcription. The team has implemented lightweight shell and [Python 'glue' scripts](#) to parse the XML outputs, extract plain text and compute simple statistics on XML elements across directories. An [ALTO viewer](#) developed by a student was adapted to browse and review the data. This viewer maps XML to the JPEG page image and enables manual editing of the ALTO files.

So far, the tests have shown that language detection and perplexity thresholds reliably filter well-extracted text from low-quality optical character recognition (OCR) outputs, enabling automated downstream processing. To this end, the team plans to employ the off-the-shelf NER tool NameTag3, as described in the [original workflow](#). Experimental results using Czech and multilingual NameTag3 models suggest that it is possible to adapt language-specific NER tools and fine-tune them using domain-specific CoNLL-U-style datasets (UDPipe-derived parses with entity annotations specific to the domain). This would allow for both structured metadata extraction (geographic names, artefact types and numeric identifiers) and vocabulary-driven entity extraction. Trials with the LINDAT Keyword Extraction tool were also conducted at the end of 2025. The analysis and evaluation of these initial results, as well as outlining the final pipeline, will be completed in 2026, with the aim of achieving a production-ready solution by 2027. Based on the results, the possibility of using automated translations with tools developed in T3.4 will be considered.

In parallel, the ARUP/B team worked on a complex update to the aggregation pipeline for AMCR data in the ARIADNE Portal. Two team members attended a workshop in Crete in 2024. Based on this experience, they used the latest version of the 3M mapping tool to completely rework the original mapping used in the aggregation process. In collaboration with ADS, the new version was tested in the staging version of the portal

and fine-tuned, providing important feedback for the portal developers. At the end of 2025, the entire AMCR collection, comprising over 400,000 records relevant for ARIADNE, was [published on the public portal](#). Furthermore, a weekly update using the AMCR API will be set up in early 2026 to keep the data continuously aligned with the original repository. The mapping will automatically employ new metadata resulting from the enrichment processes when available and stored in the AMCR. However, there may be a need for slight updates to the final version. This step will entail only a minor task in the context of the renewed aggregation pipeline for Czech data.

Next steps

Similar to T5.1.1, the enriched Czech data will become part of the AMCR repository datasets and will be automatically ingested into its native interfaces, as well as into the ARIADNE Graph DB and Portal. Even during the pilot phase, the team aims to work with extensive collections that represent the majority of the data available in the repository. Once the production workflow is ready, automated enrichment processes will be applied to all suitable data within the AMCR repository.

2.2.4 Information extraction and enrichment from speech-based data

In this task, the team aims to systematically extract named entities from unstructured, transcribed text in archaeological context sheets, thus enriching corresponding database records with machine-readable metadata. The process involves receiving Automatic Speech Recognition (ASR) transcribed text from T5.4 as input and identifying named entities related to the archaeology domain using domain-specific Named Entity Recognition (NER) models. A [preliminary workflow](#) based on this work has been uploaded and is under review at the [Social Sciences & Humanities Open Marketplace](#). For specific implementation details, the source code is available in the ATHENA RC [atrium-csd-ner](#) GitHub repository.

The current development is based on a set of 584 context sheet descriptions provided by the consortium that were initially annotated using an Large Language Model (LLM, Claude 3.7 Sonnet) based on definitions from a labelset similar to a [schema](#) proposed by A. Brandsen. To advance this work, the team sought advice on the labelset by consortium experts to ensure the semantic categories are appropriate for the domain. Consortium members will use this labelset and a [graphical interface](#) to correct automatic annotations on a subset of the context sheet descriptions. The correction of these automatic annotations will help develop a small development and evaluation dataset, allowing for the refinement of domain-specific models. See Figure 2.10 for the output of the current model.



Atrium Speech-based NER demo

Text to process:

This context is about exploring the depth of the possible ditch. Lots of Late Neolithic II pottery, and some pieces of LNI. Also, shells, a few bones, one of which is burnt, ground stones and a stone tool.

Process text with NER

Results:

This context is about exploring the depth of the possible ditch **CONTEXT** .

Lots of Late Neolithic II **PERIOD** pottery **ARTIFACT** , and some pieces of LNI **ARTIFACT** .

Also, shells **ARTIFACT** , a few bones **ARTIFACT** , one of which is burnt, ground stones **ARTIFACT** and a stone tool **ARTIFACT** .

Figure 2.10: ATRIUM Speech-based NER demo

2.3 Process knowledge graph demonstrator (T5.1.3)

Summary

A prototype demonstrator was developed to showcase how a work-process knowledge graph can enhance retrieval, exploration, and visualisation of complex research processes, implemented in Neo4j and GraphDB.

Published Workflow

<https://marketplace.sshopencloud.eu/workflow/tZ6360>

2.3.1 Overview

Task 5.1.3 showcases the application and potential uses of a work process Knowledge Graph, created from the results of T4.1.3, by developing a prototype Knowledge Base and a series of targeted use case scenarios. Specifically, the team demonstrates the retrieval capabilities via relevant queries, highlighting the types of complex questions the knowledge graph can address. The demonstrator will also include a user-friendly interface for faceted exploration of the knowledge graph. Additionally, the team will present visualisations that provide an overview of all work processes, facilitating their exploration, classification, and analysis – such as detecting communities through clustering or identifying influential elements based on centrality measures.

The work carried out so far has focused on establishing the technical backbone for the knowledge graph, validating interoperability across different graph technologies ([Neo4j](#) and [GraphDB](#)), and preparing the foundation for integrating datasets into the system. The demonstrator will ultimately illustrate how knowledge graph technologies can enhance the retrieval, exploration, and visualisation of complex research processes. A local Neo4j instance has been successfully deployed and tested. In order to test the functionalities of the demonstrator, an initial manually-curated dataset with 1000 sentences from publications in archeology was inserted as input for the Neo4j Knowledge Base. The configuration and functionality were validated through a series of indicative queries, whose code in the corresponding query language (Cypher), natural language equivalents, and results are summarised [here](#). In parallel, Python scripts were developed in order to automatically convert the knowledge graph into RDF format, ensuring interoperability with [GraphDB](#) and supporting users who prefer working in an RDF environment. This dual compatibility offers flexibility while maintaining a unified data model across tools. The RDF version of the dataset can be found [here](#).

Parts of this work were presented at the [CAA conference](#) (2025) (pp. 314–316), allowing the team to showcase their work and gather feedback. The RDF version of the data was loaded on a dedicated repository of the ARIADNE GraphDB instance. SPARQL queries have been defined to respond to the same series of questions run on Neo4j to verify alignment and query expressiveness. The set of queries are available on [GitHub](#).

2.3.2 Provision of sample datasets

At this stage, both Knowledge Base (KB) instances (Neo4j and GraphDB) have been populated with a sample dataset derived from [JSTOR](#). This manually-curated dataset served as a first step in order to create a prototype demonstrator and test technical functionalities such as import methods, data conversion scripts etc. for both technologies (Neo4j and GraphDB). Following this initial phase, the team will shift focus towards open-access sources, particularly the [ADS repository](#). Integration of additional curated datasets from selected journals and reports from ADS will enable the creation of additional knowledge bases with dedicated query sets, allowing users to fully explore the structure, relationships, and semantic richness of the produced knowledge graphs. ADS is also one of the contributors of the ARIADNE Knowledge Base and it can be therefore used as a pilot to bridge with the ARIADNE ontology and enrich the ARIADNE Knowledge Base with additional information about research processes.

2.3.3 Knowledge graph development

Neo4j Knowledge Base Demonstrator

The Neo4j Knowledge Base (KB) Demonstrator utilises the [Neo4j](#) Community-edition Graph Database and is currently installed on a local environment. Access to the KB and CRUD operations are supported both visually – through a graphical user interface – and programmatically via a Cypher Endpoint. Retrieval capabilities are supported in cypher language either through prefixed queries that appear as dedicated buttons in the UI (Figure 2.11), or via execution of ad-hoc queries in Cypher language, from the corresponding input field in the visual user interface (Figure 2.12). The schema of the knowledge graph, based on which the queries are executed, is derived from [Scholarly Ontology](#) (SO), a conceptual framework for modeling scholarly work. Figure 2.13 shows the schema in the current implementation of the Neo4j demonstrator. Apart from the semantic nodes and relationships that are supported by SO, additional properties have been created based on mappings to the corresponding publication metadata schema of the available JSTOR dataset. These can be adjusted accordingly, in order to fit other publication metadata schemas (e.g. ADS).

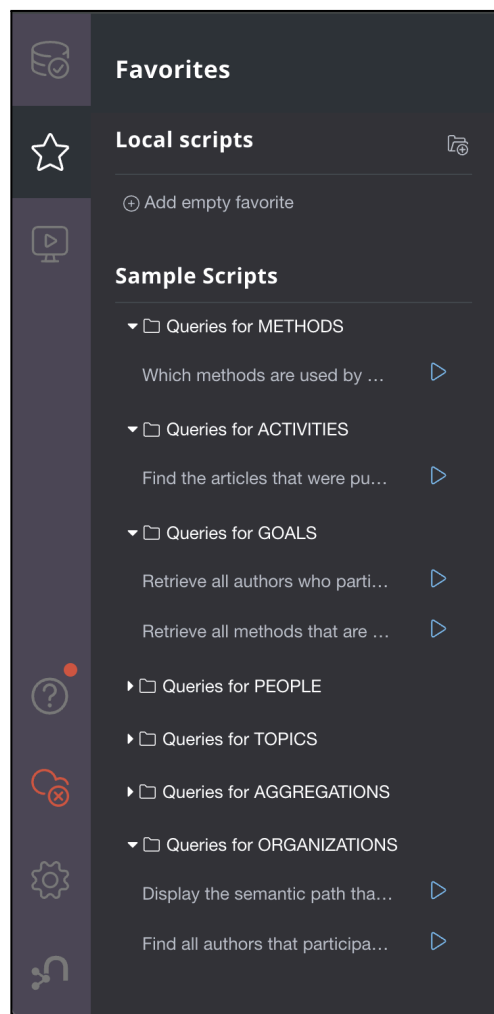


Figure 2.11: Neo4j prefixed queries interface



Figure 2.12: Neo4j query input field.

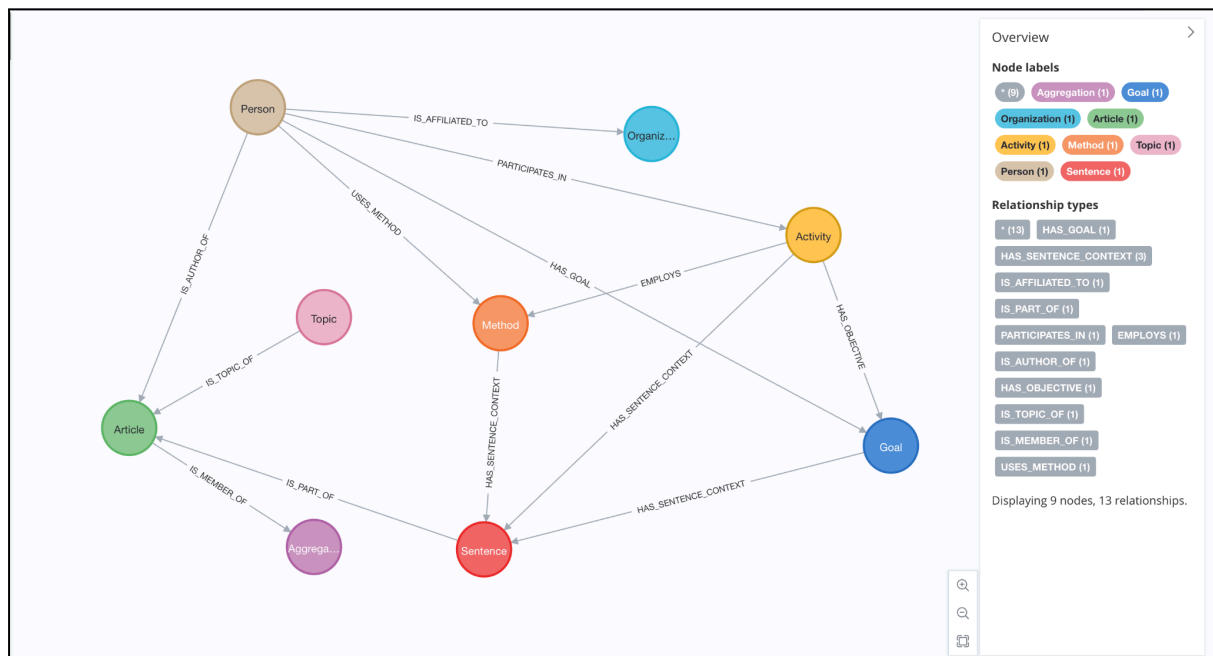


Figure 2.13: Neo4j KB schema.

The retrieval capabilities of the demonstrator are presented through queries in cypher language that exploit the semantic paths of the graph and offer results in tabular, code, visual or textual form, depending on the output. Indicative query examples, presented in different output forms are displayed below (Figures 2.14–2.17):

```

1 // Retrieve all methods that are used by activities which have objectives related to
  "analysis" or "date".
2
3 MATCH (g:Goal)-[:HAS_OBJECTIVE]-(a:Activity)-[:EMPLOYS]-(m:Method)
4 WHERE g.textual_span CONTAINS 'analysis' OR g.textual_span CONTAINS 'date'
5 RETURN DISTINCT m.name AS MethodName
6 ORDER BY m.name

```

MethodName
"Detrended correspondence analysis (DCA)"
"Redundancy analysis (RDA)"
"Tuff I group method"
"archaeological and dendrochronological analyses"
"linear regression"
"trace element analysis"

Started streaming 6 records after 1 ms and completed after 4 ms.

Figure 2.14: Query result in table form.

```

neo4j$
1 // Retrieve all authors who participate in activities employing specific methods and
  whose objectives relate to "reconstruct" or "vegetation" or "environment". For each author,
  show the methods used and the corresponding goal.
2
3 MATCH (p:Person)-[:PARTICIPATES_IN]-(a:Activity)-[:EMPLOYS]-(m:Method)
4 MATCH (a)-[:HAS_OBJECTIVE]-(g:Goal)
5 WHERE g.textual_span CONTAINS 'reconstruct'
6 AND (g.textual_span CONTAINS 'environment' OR g.textual_span CONTAINS 'vegetation')
7 RETURN DISTINCT p.full_name AS Author,
8 COLLECT(DISTINCT m.name) AS Methods, g.textual_span AS Goal
9 ORDER BY Author

```

Author	Methods	Goal
"Aaron P. Potito"	["Pollen analysis"]	"reconstruct vegetation"
"Arghya K. Hait"	["sediment and pollen analysis"]	"reconstruct the Holocene mangrove and environmental changes at a coastal site Pakhira in the Sundarban Biosphere Reserve in the western Ganga-Brahmaputra Delta, India"
"Carole Adolf"	["charcoal analysis"]	"reconstruct the vegetation and fire history in north-eastern Sardinia"
"Carole Adolf"	["pollen and spore analysis"]	"reconstruct extra-local to regional vegetation dynamics"
"Carole Adolf"	["pollen and spore analysis"]	"reconstruct local vegetation dynamics"
"Daniele Colombaroli"	["charcoal analysis"]	"reconstruct the vegetation and fire history in north-eastern Sardinia"
"Daniele Colombaroli"	["pollen and spore analysis"]	"reconstruct extra-local to regional vegetation dynamics"

MAX COLUMN WIDTH: 100

Figure 2.15: Query result in textual form.

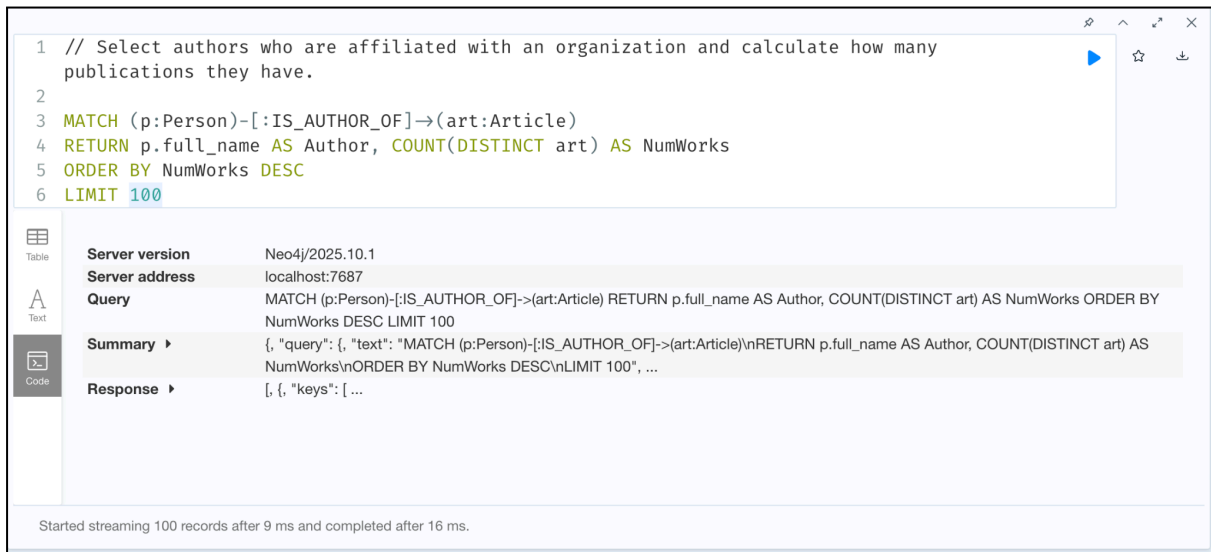


Figure 2.16: Query result in code form.

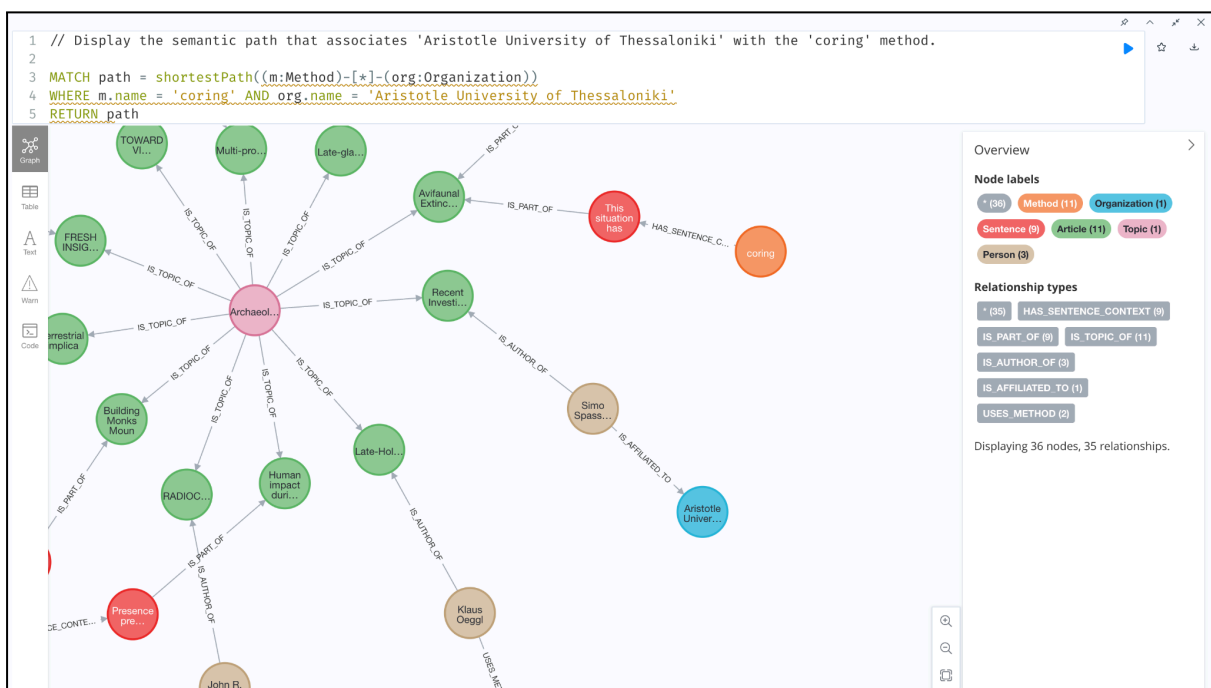


Figure 2.17: Query result displayed in graph form.

GraphDB Knowledge Base Demonstrator

The GraphDB Knowledge Base (KB) Demonstrator utilises the GraphDB instance of ARIADNE at the [D4Science.org](https://d4science.org) infrastructure hosted by Consiglio Nazionale delle Ricerche (CNR-ISTI). In this phase of the demonstrator, a repository of the staging instance of GraphDB, accessible only by managers of the ARIADNE infrastructure, has been used. We plan to move the demonstrator to the publicly accessible GraphDB

instance, possibly integrating the dataset in the same GraphDB repository of the ARIADNE Knowledge Base. All materials: input and SPARQL queries are available in the [dedicated GitHub repository](#).

GraphDB supports access and queries both visually – through a graphical user interface called GraphDB Workbench (Figure 2.16) – and programmatically via a SPARQL endpoint.

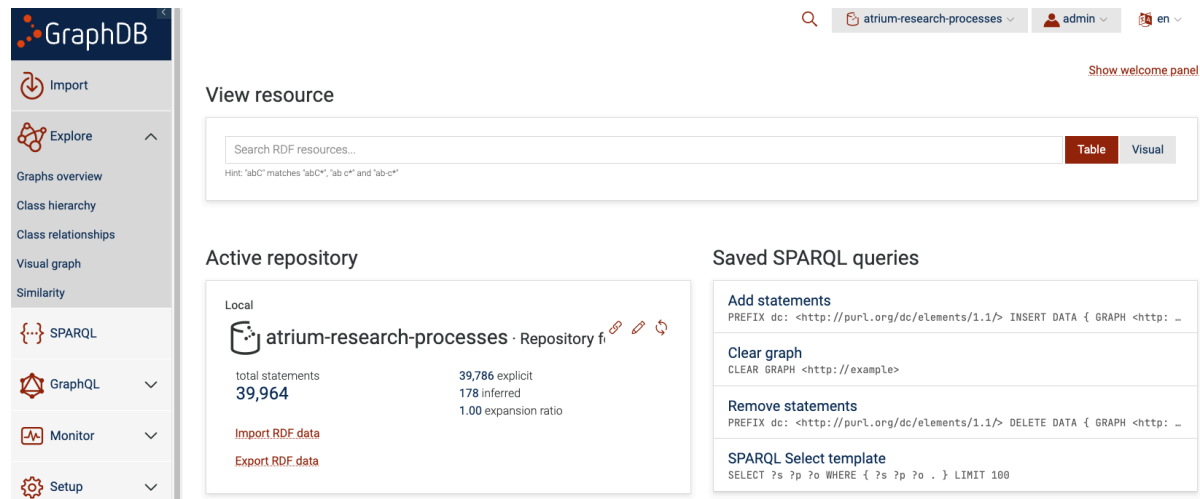


Figure 2.18: GraphDB Workbench

The repository was populated by importing the RDF version of the dataset that can be found [here](#). The retrieval capabilities of the demonstrator are presented through queries in SPARQL language that exploit the semantic paths of the graph and offer results in json or visual mode as simple table, configurable pivot table, or charts. The available options may vary depending on the type of query. In any case, results can be downloaded in different formats (json, xml, csv, tsv, rdf). Indicative query examples, presented in different output forms are displayed below (Figures 2.17–2.20):

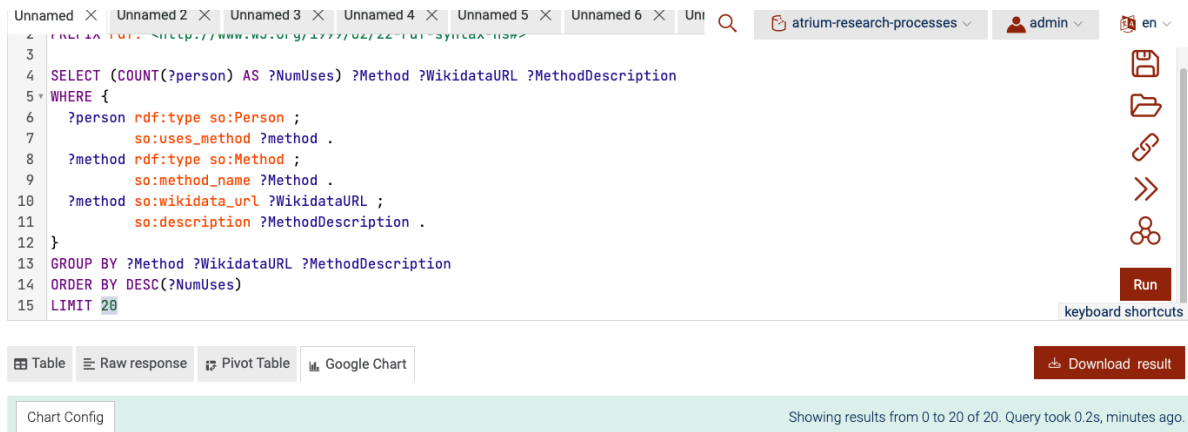
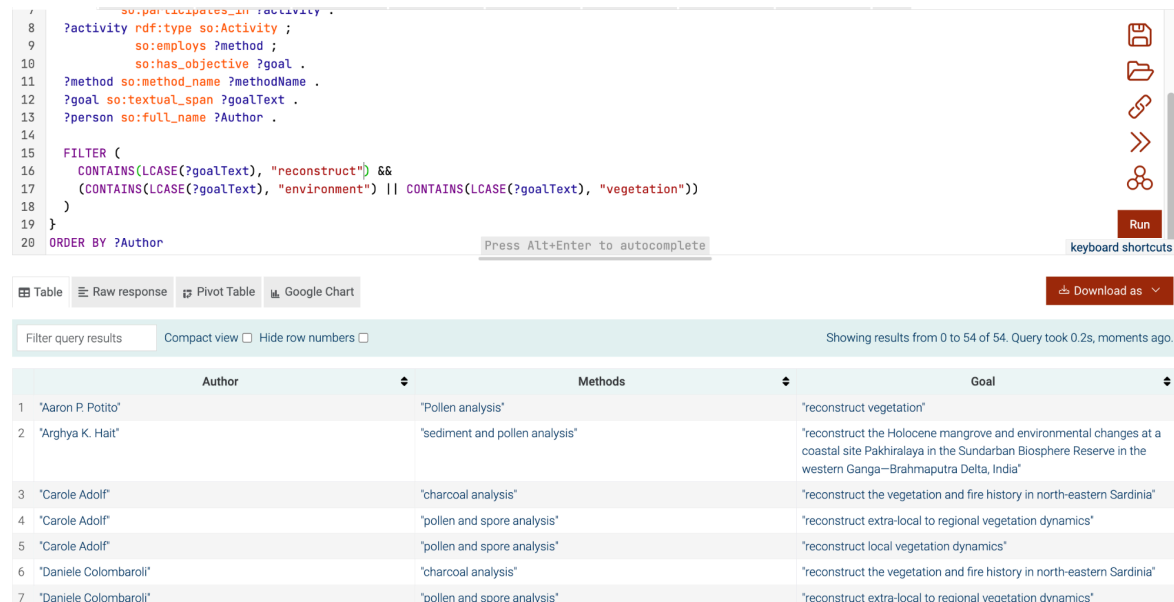


Figure 2.19: Query results in bar chart form.



The screenshot shows the ATRIUM query interface. At the top, there are tabs for 'Unnamed 2' through 'Unnamed 6'. A search bar contains 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'. Below the search bar, a SPARQL query is displayed:

```

8
9 ?activity rdf:type so:Activity ;
10          so:employs ?method ;
11          so:has_objective ?goal .
12 ?method so:method_name ?methodName .
13 ?goal so:textual_span ?goalText .
14 ?person so:full_name ?Author .
15
16 FILTER (
17   CONTAINS(LCASE(?goalText), "reconstruct") &&
18   (CONTAINS(LCASE(?goalText), "environment") || CONTAINS(LCASE(?goalText), "vegetation"))
19 )
20 ORDER BY ?Author

```

On the right side, there are icons for saving, opening, linking, and running the query. A 'Run' button is at the bottom right. Below the query, there are tabs for 'Table', 'Raw response', 'Pivot Table', and 'Google Chart'. A 'Download as' button is on the right. Below these tabs, a 'Filter query results' button is on the left, and a status bar on the right says 'Showing results from 0 to 54 of 54. Query took 0.2s, moments ago.'

The main area displays a simple tabular form with the following columns: Author, Methods, and Goal. The results are as follows:

Author	Methods	Goal
"Aaron P. Potito"	"Pollen analysis"	"reconstruct vegetation"
"Arghya K. Hait"	"sediment and pollen analysis"	"reconstruct the Holocene mangrove and environmental changes at a coastal site Pakhralaya in the Sundarban Biosphere Reserve in the western Ganga-Brahmaputra Delta, India"
"Carole Adolf"	"charcoal analysis"	"reconstruct the vegetation and fire history in north-eastern Sardinia"
"Carole Adolf"	"pollen and spore analysis"	"reconstruct extra-local to regional vegetation dynamics"
"Carole Adolf"	"pollen and spore analysis"	"reconstruct local vegetation dynamics"
"Daniele Colombaroli"	"charcoal analysis"	"reconstruct the vegetation and fire history in north-eastern Sardinia"
"Daniele Colombaroli"	"pollen and spore analysis"	"reconstruct extra-local to regional vegetation dynamics"

Figure 2.20: Query results in simple tabular form.

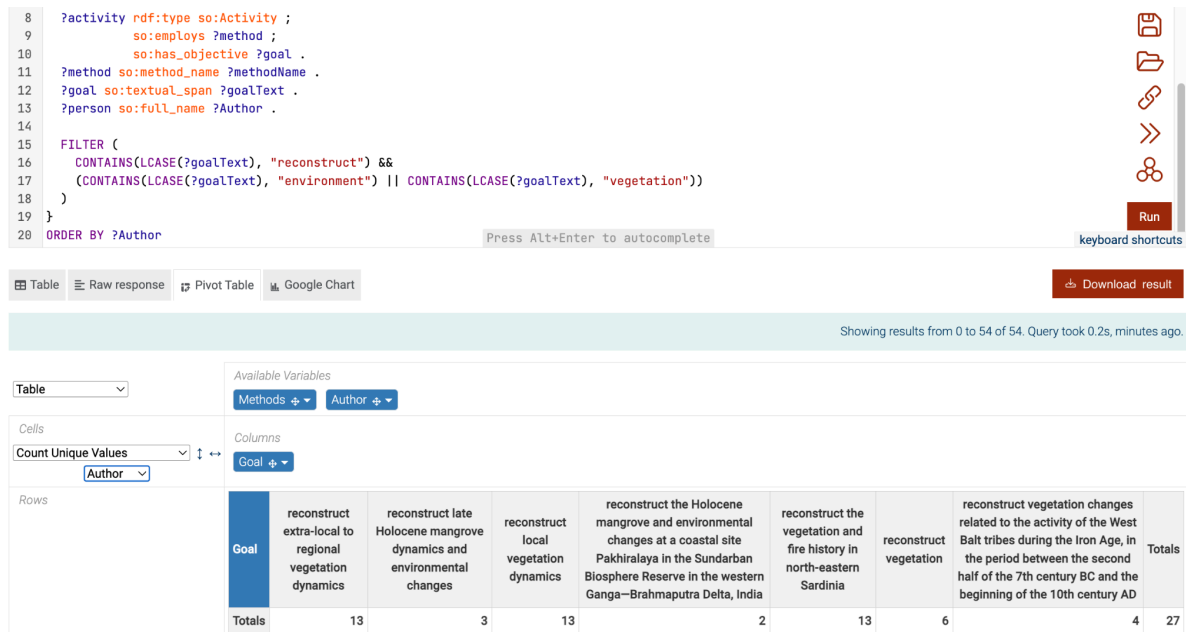


Figure 2.21: Query results in a pivot table.

It shows how many unique authors are participating in activities with a given goal.

Visual graph



Figure 2.22: Query results displayed in visual graph form

2.3.4 Provision of final datasets

Regarding ADS final datasets, the team has already exposed the [link to the PDF](#) for archaeological reports in the API so they can easily be downloaded. For journal articles, ADS has exported a static XML file containing metadata about the articles and the link to the PDFs, so all data are accessible to the T5.1.3 team.

2.3.5 Compliance with original plan

The work completed so far aligns with the original objectives defined for Task 5.1.3. The Neo4j implementation of the demonstrator provides advanced flexibility for prototyping and visual exploration as well as advanced functionalities such as the application of graph analysis algorithms, the inclusion of a vector store, etc. In addition, the RDF KG and its KB implementation in GraphDB ensures compatibility with SPARQL endpoint and Semantic Web technologies. This dual approach provides future users with the freedom to adopt the graph format best suited to their workflows and ensures that the demonstrator remains aligned with open and [FAIR](#) data principles.

2.3.6 Next steps

Over the remaining two years (M25–M45), the team will:

- Develop additional ready-to-use Neo4j and SPARQL queries to support specific research questions.
- Adapt the Knowledge graph schema in order to be compatible with publication metadata fields of the ADS.
- Create a demo dataset from an indicative selection of ADS articles (journals and reports) that will function as a separate source for the Knowledge Graph.
- Create additional Neo4j and GraphDB knowledge bases along with indicative queries tailored for the ADS dataset.
- Deploy and host the demonstrator online.

The expected timeframe for the deployment of the full version of the demonstrator is late 2026 to early 2027.

3. T5.2 Image-Based Demonstrators

Introduction

Archaeological data is very image-rich, whether these are photographs of individual artefacts, monuments or parts of monuments. Archaeologists often need to compare and annotate 2D images. Recently there has been considerable interest within the Arts and Humanities in the [International Image Interoperability Framework \(IIIF\)](#). IIIF viewers allow researchers to view images drawn from multiple distributed repositories, and to annotate images, creating their own curated collections. Photography is also widely used in archaeological site and artefact recording, and repositories often hold large archives of photographic images but these often lack rich metadata. In the UK and Czechia there are national databases of images of finds made by members of the public. In response to a need identified in T2.3.1 from the UK and Czech metal-detector communities to have online databases of classified finds to serve as reference collections, the team therefore wanted to explore whether AI approaches could be used to identify finds from 2D images. The two image-based demonstrators therefore cover two use cases: the first demonstrator uses the workflow developed in T4.2.1 to implement the IIIF viewer at ADS and in the ARIADNE portal allowing the online comparison and manual annotation of research datasets (T5.2.1 and 5.2.2), while the third demonstrator is more experimental as it is intended to showcase the workflow developed in T4.2.2 for AI annotation of archival photographic collections and photographs of metal-detected finds.

Work to date on the image-based demonstrators has delivered Altamira, a forked Mirador-based IIIF viewer with enhanced Web Annotation Data Model compliant annotation capabilities, tested on a dedicated server and deployed in staging environments for ADS and ARIADNE, together with pyramidal image generation and IIIF manifests for key collections such as Stained Glass and Bronze Age Rock Art, interface updates on the ADS and ARIADNE portals. On the Czech side, work so far has focused on the AI-based image demonstrator, assembling multiple object and archival photo datasets, producing an annotated AMCR-PAS training set, and developing an initial generation of segmentation and classification models alongside an in-progress SKOS vocabulary mapping between PAS and AMCR-PAS to underpin subsequent re-training and integration into the AMCR Digital Archive and the ARIADNE Portal.

Workflows

- <https://marketplace.sshopencloud.eu/workflow/YHAnKE>
- <https://marketplace.sshopencloud.eu/workflow/6NGfcj>

- <https://marketplace.sshopencloud.eu/workflow/G6ck4w>

3.1 Altamira, Mirador-based IIIF viewer

Altamira is a fork of Mirador, the widely adopted IIIF viewer. The software has been re-engineered to align with specific project requirements, with the most notable enhancement being the integration of robust annotation support. To ensure maximum interoperability, these annotations adhere to the Web Annotation Data Model (WADM) standard; this allows data to be exported and reused across different instances of Altamira or any external software supporting WADM.

Functionality is further enhanced by workspace persistence capabilities. Users may preserve their distinct workspace environments via browser local storage or export them as files for subsequent use. These features are designed to streamline the research workflow, facilitating knowledge sharing and data exchange between scholars. The viewer ensures cross-browser compatibility and supports comparative analysis by enabling users to view multiple resources simultaneously in customisable configurations (e.g. side-by-side, grid, or columnar layouts).

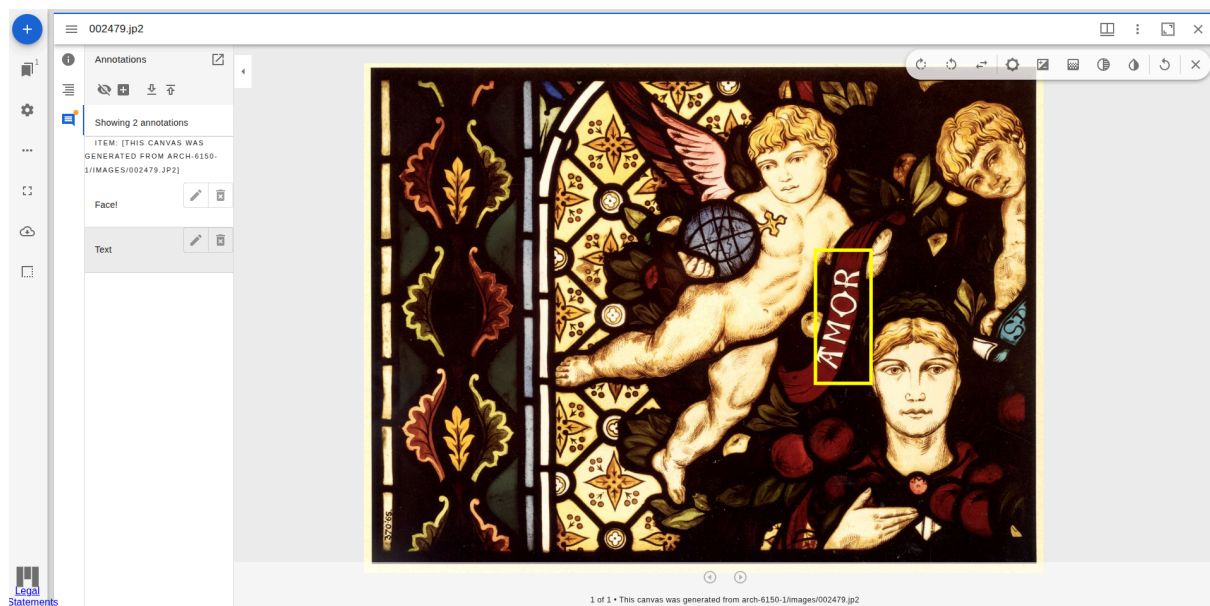


Figure 3.1: Altamira IIIF Viewer showing annotations on stained glass.


Adhering to the open-source philosophy of its predecessor, [Altamira](#) remains fully open source. This transparency allows institutions to self-host or modify the source code to meet specific architectural needs. Altamira's operational readiness and practical utility

have been demonstrated through its successful deployment in staging environments for the ADS and ARIADNE projects.

While developing Altamira, PCSS maintained a test server. There were two different rounds of thorough testing to ensure the quality of the product. The team also prepared a branded version of Altamira for ADS to display a cookie banner and the university's Impressum.

3.2 Interface updates at the Archaeology Data Service

Based on the Altamira deployment software, ADS had generated pyramidal images and manifests, which are both needed to use a IIIF viewer. The next step will be to enrich the manifests with metadata concerning the visual resources so users have a better idea of the ownership, license, and context of the resources, should they access the viewer directly without passing through the ADS website or the ARIADNE portal.

After collecting information about potential user flows and needs ( IIIF Users), a solution was designed to enable users to collect IIIF manifests on the ADS website interface so they can view and annotate multiple resources in the Altamira viewer. Users can now open a visual resource in Altamira from the landing page of the resource or download the manifest for the [Rock Art collection](#), and Stained Glass will follow early in 2026.

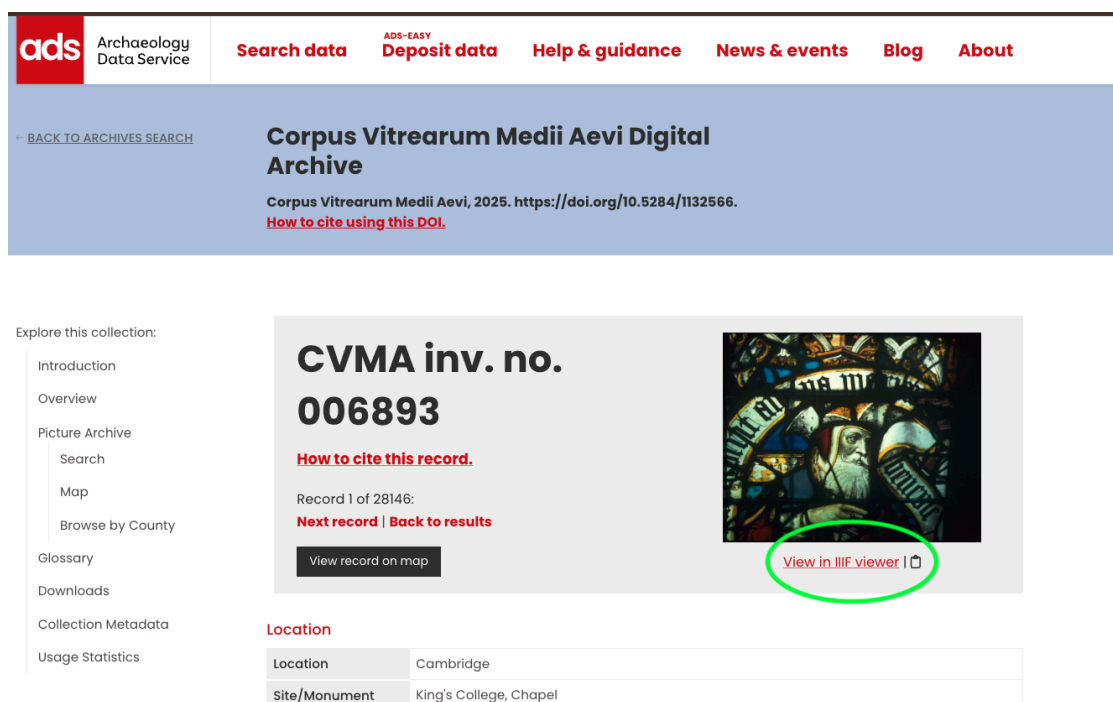
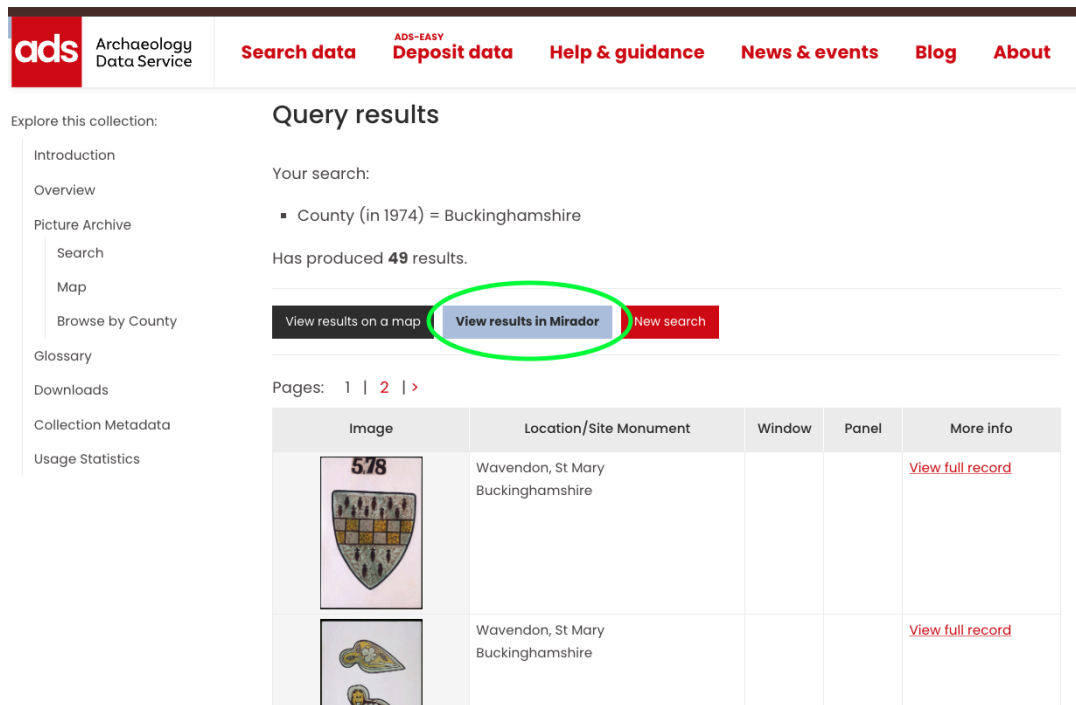


Figure 3.2: ADS interface showing the "View in IIIF viewer" button.

If search results are fewer than 150, it is possible to download a workspace that includes all of the manifests to import all these resources more easily into Altamira.



ads Archaeology Data Service

Search data **Deposit data** **Help & guidance** **News & events** **Blog** **About**

Explore this collection:

- Introduction
- Overview
- Picture Archive
 - Search
 - Map
 - Browse by County
- Glossary
- Downloads
- Collection Metadata
- Usage Statistics

Query results

Your search:

- County (in 1974) = Buckinghamshire

Has produced **49** results.

View results on a map **View results in Mirador** New search

Pages: 1 | **2** | >



Image	Location/Site Monument	Window	Panel	More info
	Wavendon, St Mary Buckinghamshire			View full record
	Wavendon, St Mary Buckinghamshire			View full record

Figure 3.3: ADS query results showing "View results in Mirador" option.

The team is currently working on documentation for the users of the ADS interface to improve usability. Changes were also made in the ARIADNE portal to facilitate usage: these changes are described below in Section 8.

3.3 Early mediaeval sculpture (T5.2.1)

In task 5.2, two case studies were chosen (T5.2.1 & T 5.2.2) to illustrate the [associated workflow](#) T4.2.1: "IIIF". These case studies were selected to showcase single discipline and multi-disciplinary research usage of the IIIF standard. Early Medieval Sculpture of England has been catalogued in a [major corpus](#) project funded by the British Academy over the last 40 years. The motifs are drawn from rich iconography of the period which can also be seen in illuminated manuscript art. The digital corpus is currently being made available online by the ADS. The aim of the demonstrator will allow researchers to compare specific pieces of stone sculpture from the ADS repository with illuminated manuscripts which are made freely available online by the British Library and other major European Libraries and Archives.

Due to unforeseen issues regarding image reuse permissions, the team was not able to use the early mediaeval sculpture dataset as the first example as planned. Those issues are now resolved, so the team expects to work on this dataset in the first quarter of 2026. In the meantime, the focus pivoted to use the Stained Glass collection, which also provided rich potential in terms of using image annotations for comparative research for the [collection](#). As described above, handy links were added in the ADS interface to provide users with easy access to viewing and annotating the resources using the Altamira viewer. The [dataset](#) was also imported into the ARIADNE portal for further discoverability. The collection was organised in ARIADNE into a tree-like structure to facilitate navigation. For instance, the main collection is the parent to church records, which are parents to panel records, which contain windows as children. The metadata of the records was enriched by mapping a large proportion of the native subject terms to the Getty AAT, enabling these stained glass windows to be found by the motifs they display, in concomitance with other artifacts displaying similar motifs. From this experience with the stained glass collection, the team will replicate the process with the stone sculpture collection, which will be available in the same way, from the ADS special collection page and from the ARIADNE portal.

3.4 Bronze Age Rock Art (T5.2.2)

For the second demonstrator of the IIIF viewer, the team decided to focus on a single-discipline case study, with sources drawn from two countries, the UK and Sweden, both represented in ATRIUM partner data providers. Thought to largely derive from the European Bronze Age, Rock Art is prevalent on rock surfaces in Northern England and Sweden. The motifs comprise a rich variety of geometric and naturalistic depictions. There are national databases of rock art images for Northern England and Sweden held by ADS and SND-SHFA respectively. However, it is difficult to compare the motifs from Britain and Sweden which would allow researchers to identify areas of commonality and difference and potentially determine if they have shared or parallel trajectories over time. The recording of rock art has often been undertaken by groups of volunteers, using paper rubbings, photographs, and increasingly photogrammetry, linking this task to the community groups identified in T2.2.3.

ADS

The collection of [Bronze Age Rock Art at ADS](#) is one of the two datasets used in this task. After a special collection interface for this dataset was prepared, the team exported an XML output to be ingested in the ARIADNE portal. This included enriched metadata, including motif types as native subjects, along with links to IIIF manifests. To perfect this dataset, the team spent efforts in curating a representative selection of

images for each record for the ARIADNE portal. The collection can be found in the [ARIADNE portal](#).

SND

The second dataset part of this task is the Swedish Rock Art Research Archives ([Svenskt Hällristningsforskningsarkiv](#), SHFA), archived by SND. A new OAI-PMH API was designed so that records can be fetched automatically and updated seamlessly in the ARIADNE portal to keep the dataset up to date. Improvements have been made to the metadata, images thumbnails were deployed, and IIIF manifest links were added so they are accessible in the ARIADNE portal. The collection can be found in the [ARIADNE portal](#) but will be updated again early in the new year to display the newly provided IIIF links through the OAI-PMH API.

When the demonstrator is completed, during 2026, it will be possible to combine manifests from these two sources (ADS and SND-SHFA), to annotate them in Altamira, and to preserve and share these annotations with colleagues for research purposes.

3.5 Archival photographic collections image annotation (T5.2.3)

3.5.1 Overview

This task demonstrates how ML-models for semi-automatic image recognition and metadata extraction can be applied on archaeological image archives, with a particular focus on the archival photographic collections of ARUB and ARUP. The models are developed on sample datasets and a selection of AMCR controlled vocabularies using the [associated workflow](#) (task T4.2.2), with envisaged outputs including annotated training datasets, controlled vocabulary mappings creating links between datasets, trained ML-models, and metadata records published via the AMCR Digital Archive and aggregated by the ARIADNE Portal.

So far, most effort has focused on developing a computer vision ML-model capable of recognising object types in photographs, including the segmentation and tagging of the AMCR Portable Antiquities Scheme (PAS) dataset and the ongoing mapping of controlled vocabulary terms between AMCR-PAS and British PAS using SKOS mapping properties.

On top of what was planned, short textual descriptions will be created for each image using LLMs and a visual transformer will be used to create image embeddings to analyse similarity between images and allow similarity search. Both these tasks were

experimented with at the beginning of the work and their implementation seems plausible.

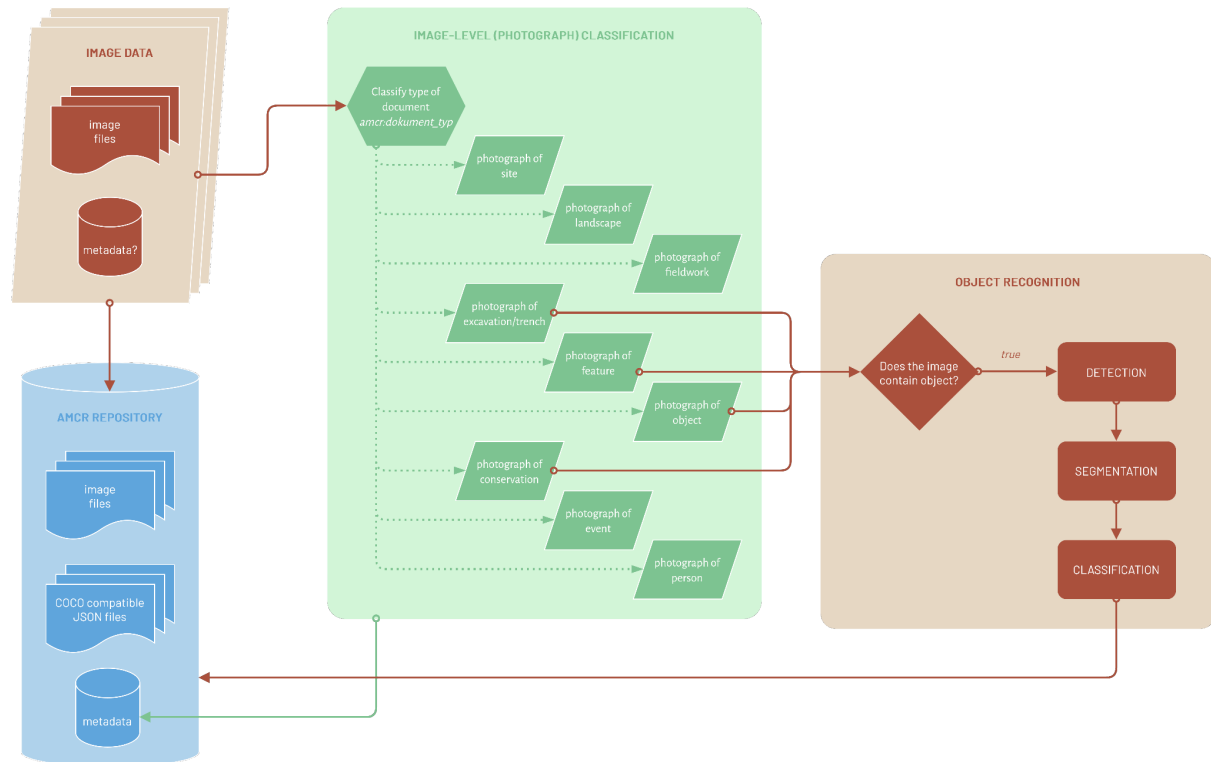


Figure 3.4: The pipeline a photograph will go through when entering the AMCR system

3.5.2 Provision of datasets

At this stage, sample datasets serve mostly as training data for fine-tuning computer vision models. Two types of sample image datasets were collected: (1) photographs of archaeological artefacts (objects), mostly collected by metal detecting communities, and (2) archival photographs of fieldwork, sites, excavations etc.

1. Object datasets:

- [AMCR-PAS dataset](#) (provided by ARUP and ARUB), comprising more than 7,500 metal detecting finds with ca. 10,000 photographs and associated metadata (harvested from [OAI-PMH API](#) and a file serving API);
- [PAS dataset](#) (Portable Antiquities Scheme, provided by UoY-ADS) consisting of ca. 750,000 mostly metal detecting finds with images and metadata, harvested using ARIADNE Knowledge Base/ARIADNE SPARQL Endpoint and PAS website itself;
- [Lovec Pokladů dataset](#) (negotiated by ARUP and ARUB and provided by a third party) with over 300,000 finds by metal detectorists and amateur archaeologists;

- *Montelius dataset* (negotiated by ARUP and ARUB, provided by [Naturhistorisches Museum Wien](https://www.nhm-wien.ac.at/) (Natural History Museum Vienna) comprising various photographs, including mostly photographs of objects and sites, and drawings of objects.



Figure 3.5: Typical examples of metal detecting finds from AMCR-PAS: fibula (<https://doi.org/10.71928/M-202300087-N00394>), coin (<https://doi.org/10.71928/C-202109132-N00447>) and an arrowhead (<https://doi.org/10.71928/M-202101534-N00591>)

2. Archival photographs datasets:

- [Digital photographs of fieldwork collection](#) provided by ARUP and available through AMCR APIs, consisting of around 60,000 born-digital photographs of various scenes;
- *Moravika legacy photographs collection* (provided by ARUB) of around 2,000 digitised black and white photographs of excavations, fieldwork, sites and objects;
- *Dolní Věstonice and Pavlov collections* of around 8,000 digitised negative photographs (provided by ARUB) comprising photographs of excavation, fieldwork, sites, objects etc.

3.5.3 Outputs

At this stage, the main output of the sub-task is an annotated AMCR-PAS dataset consisting of around 10,700 of annotations, i.e. polygon segments identifying the exact outline of the artefact in 9,700 photographs and associated labels. There are 112 classes from the AMCR object classes controlled vocabulary represented in the dataset. The dataset is stored as a combination of image files in JPG and associated Common Objects in Context (COCO)-compatible JSON file(s) with annotations.

Another major output will be a mapping between controlled vocabularies employed by PAS and AMCR-PAS using SKOS properties which is currently under development. This

will allow re-training the model on a larger training dataset and hence improve its quality.

Last but not least, the first generation of segmentation and classification ML-models is also available, but due to limited amount of training data (AMCR-PAS dataset) and high imbalance in the dataset, it has some limitations. The segmentation model performs sufficiently well to reduce human input needed for creating polygons in the photographs of finds and will significantly reduce the time needed for further annotations.



Figure 3.6: An example of semi-annotated archival photograph – documentation photograph of an Early-Medieval grave with skull and bones, sword, spurs and several stones segmented using polygons (original image: <https://doi.org/10.60585/M-FT-110736000>)

4. T5.3 3D-Based Demonstrators

Introduction

With the growing affordability of laser scanning and other forms of 3D recording, archaeologists are making increasing use of 3D data in artefact and site recording. As a result, better research workflows and accompanying processing and dissemination pipelines are needed to streamline working with resulting data outputs. The workflows and viewers developed in T4.3: “3D-Based workflows” form the basis of the 3D-based demonstrators developed in T5.3 which focus on standing buildings. Standing structures are complex and sensitive to human alteration and natural climatic impacts. The 3D demonstrators have therefore been developed to enable researchers and those involved in Cultural Heritage (CH) asset management to collaborate remotely and exchange knowledge via an interface that integrates conservation and analytical data with archaeological and historical information about building use and transformations over time, as well as the environmental and cultural context. Task 5.3 drives the development of this proposed service by creating demonstrators that represent different cultural heritage (CH) asset typologies and address different objectives (e.g. risk assessment, conservation, valorisation).

To do so, a few case studies have been identified and selected to respond to these objectives and are currently under development. The work carried out so far has focused on the preparation of the 3D models related to the case studies. Moreover, various integration and visualisation tests in the selected viewers have been undertaken. The different case studies are each at a different level of development.

The EpHEMERA platform has been expanded with new visualisation capabilities, including PoTree for point cloud visualisation and 3DHOP for textured mesh models. Three case studies in Cyprus (Ayios Ioannis Lampadistis Monastery, Kampanopetra Basilica, and Pyla-Kokkinokremos) have been prepared with 3D models at various stages of integration and testing in the selected viewers.

The next steps for the T5.3 team will be to:

- solve the issues of the 3D model visualisation in the viewers through the finalisation of testing;
- integrate the 3D models related to the case study in the selected viewer(s);
- finalise the cases studies through the development of storytelling projects;
- give access to the T5.3 Demonstrator through the aggregation of the case studies in the ARIADNE portal.

Overarching associated workflow: [T4.3_3D-Based workflows](#) .

4.1 3D architectural models (T5.3.1)

Summary

Outputs from Task 5.3.1-3D architectural models are to be used to create three demonstrators within the EpHEMERA platform to explore 3D visualisation for heritage management, conservation, and research. New visualisation tools and workflows will be implemented for three case studies in Cyprus:

- Ayios Ioannis Lampadistis Monastery;
- Kampanopetra Basilica;
- Pyla-Kokkinokremos urban settlement.

This task has developed three demonstrators within the EpHEMERA platform to explore 3D visualisation for heritage management. The team has deployed and tested multiple viewers (PoTree, 3DHOP, ATON), and customised these tools for specific heritage data needs. PoTree implementation for the Ayios Ioannis Lampadistis Monastery is now available online, and 3DHOP visualisations for the Kampanopetra Basilica have been completed.

Workflow

<https://marketplace.sshopencloud.eu/workflow/W2DT4S>

4.1.1 Provision of sample datasets

For the demonstrators of 3D architectural models, the outputs described in T5.3.1 (3D architectural models) were used as the primary input. The task aims to drive the development of the proposed service through the creation of three case studies, addressing specific challenges:

1. a monument;
2. an archaeological site;
3. a heritage building within its historic urban built environment.

These examples will explore how 3D architectural models may function as spatial interfaces that provide access to multiple discipline information through advanced viewers. Item level access to each case study will be provided via the ARIADNE portal.

Specifically, the EpHEMERA service, developed by the Cyprus Institute and already integrated within the [ARIADNE services](#), will be the point of access for the advanced viewers and the visualisation of the T5.3.1 demonstrators.

4.1.2 EpHEMERA

[EpHEMERA](#) is a platform allowing users to visualise and interact in 3D with layers of archaeological excavations, ancient buildings, archaeological areas, and their related documentation. The primary objective of the service is the documentation and management of Cultural Heritage at risk (e.g., invasive urban development, war conflicts, natural and human agents, inaccessibility). The 3D interactive online geo-service is aimed at the preservation of endangered architectural and archaeological heritage. The EpHEMERA system is intended to serve as a multidisciplinary infrastructure where it is possible to:

- Visualise online and through standard web browser 3D architectural and archaeological models classified according to a specific type of risk;
- Query the database system, thanks to a standardised metadata structure used to describe the CH asset models, and retrieve information attached to each digital object;
- Interact with the 3D model;
- Extract geometric and morphological information.

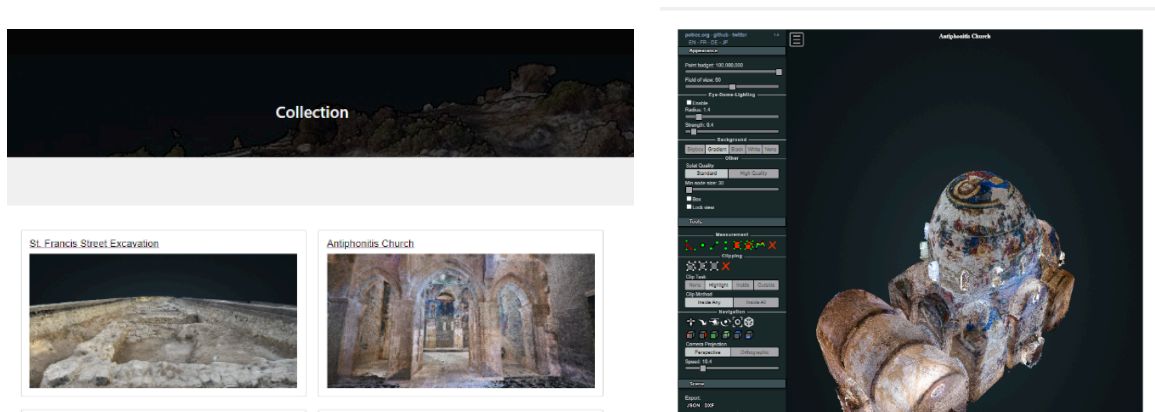


Figure 4.1. Snapshot of the EpHEMERA platform with access to: a) the whole collection and b) to the single assets.

Currently, the service addresses the 3D documentation, management and investigation of built heritage point clouds. Through the open source [PoTree viewer](#) and its tools, users can access and visualise online 3D architectural and archaeological models, extract geometric and morphological information and conduct various measurements, such as areas or volumes and extract plans and cross-sections.

Within ATRIUM, the T5.3.1 team will expand this service through the implementation of new tools and features to allow the 3D mesh visualisation of materials and textures, user input annotation, linking to Geographical Information System (GIS) information, and 3D reconstruction comparative visualisation. For this reason, the team is working on the integration of these new features by selecting and embedding further viewers in the EpHEMERA system platform.

To fulfill the needs of the demonstrator, it was necessary to improve the existing offer of 3D visualisation, and also to add new tools to cover different types of data and interaction paradigms. This selection of the viewers was informed by the needs of the case studies used for the development of the T4.3 workflows. To display the Building Information Modelling (BIM) data interactively, we reviewed the state of the art of open BIM viewers that could be integrated in an online platform. To deal with massive digitised 3D surface models, we chose to use 3DHOP, which is available for the demonstrator via the ARIADNE Portal, hosted at ADS. The T5.3.1 team also experimented with other visualisation frameworks (such as [ATON](#)) and other visualisation libraries and output technologies ([three.js](#) plus WebXR). In most cases, the viewers were modified and customised to better adapt to the needs of the specific demonstrators. This helps evaluate the cost-effectiveness of the customisation work, assessing how much a specialised, advanced viewer may help in effectively accessing noteworthy datasets, with respect to the use of stock viewers, which are still the preferred way to access large collections.

4.1.3 Case studies

Ayios Ioannis Lampadists Monastery, Cyprus

Located in the village of Kalopanagiotis in the Troodos Mountains (Cyprus), the Ayios Ioannis Lampadistis Monastery is a UNESCO-listed heritage site and a historical landmark dating back to the 11th century. The building is a remarkable example of Byzantine architecture, showcasing distinctive features, including architectural decorative elements, frescoes and ancient graffiti. Over time, the monastery has undergone many alterations that have significantly affected its form and elevated its cultural importance. To date, the monument continues to serve as a place of religious practice, pilgrimage and tourism, attracting a large number of visitors annually. Thus, on the one hand, it maintains its original function and constitutes a living monument. Although, on the other hand, it is exposed to various human-induced hazards, in addition to the natural ones.

Due to the presence of these hazards, the Ayios Ioannis Lampadistis Monastery was selected as a case study to develop a dynamic digital solution that can be actively and

constantly updated to provide critical information on conservation status during the building's lifecycle and that can be integrated into web-based platforms, providing a more readily available source of knowledge to different heritage experts and CH managers. Indeed, this case study demonstrates a dynamic 3D workflow for research, conservation and valorisation of built heritage that requires the design of comprehensive, accessible, and collaborative steps throughout its pipeline. The choice of open data standards and software is instrumental in enabling transparent practices and seamless data integration. Thus, the demonstrator needs to use dynamic 3D viewers that can be accommodated on the selected web platform (EphEMERA) and constantly updated with new data, therefore answering the need for constant monitoring of a living monument such as the Monastery. The team chose PoTree and 3DHOP as the designated 3D viewers for this case study.

The [PoTree project of the Ayios Ioannis Lampadistis Monastery](#) has been developed and is available online (Figure 4.2). The PoTree implementation utilises a geo-referenced point cloud of the entire monastery, consisting of the church, the courtyard and the surrounding structures. The PoTree environment allows users to interact with the model, creating sections and measurements, and visualising it in three sections: the roof, the church interior, and the church and the surroundings. Therefore the demonstrator allows professional users (e.g. architects, engineers, CH managers) to access the 3D point cloud, perform geometric analyses and collaborate online in the study and monitoring of the monument. Detailed tutorials (text and videos) support the users in the discovery and learning of all the features and tools provided by the viewer.

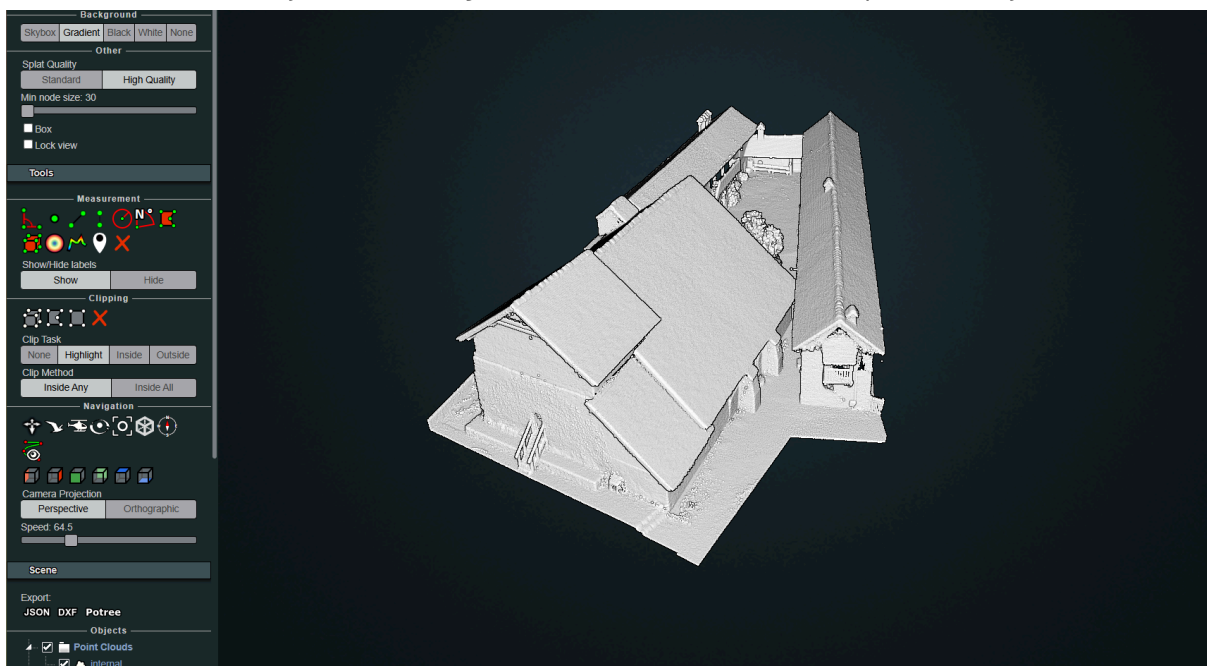


Figure 4.2. A snapshot of the interactive implementation of the monument in PoTree.

In order to expand accessibility to less technical users, the team selected 3DHOP as a second viewer. 3DHOP provides a familiar visualisation as it utilises textured meshes. Moreover, in addition to technical tools (e.g. a measuring tool), this viewer allows users to describe the cultural heritage asset, for example adding multidisciplinary information related to the whole monument as well as to its numerous components, to add annotations and links to other sources, and multiple layer visualisations. For the 3DHOP demonstrator, only the church is visualised without its immediate surroundings. Similar visualisation layers to those in PoTree were adapted in 3DHOP, allowing users to navigate inaccessible spaces of the building.

The T5.3.1 team carried out several tests for the integration of the 3D digital model in the 3DHOP viewer and, having encountered some issues, had to perform some adjustments. The first issue was related to the size of the 3D model. Indeed, the 3D digital representation is a high-resolution model resulting from the integration of different digital documentation techniques (laser scanner, photogrammetry and aerial photogrammetry). The huge size of the 3D model and the noise of some parts created several problems in the visualisation. The team had to re-process the raw data, optimising the model with simplification of the meshes and segmentation of the parts, extending the schedule for this phase of the work. After optimising the model and uploading it to the 3DHOP viewer, further issues were encountered with the textures (Figure 4.3).

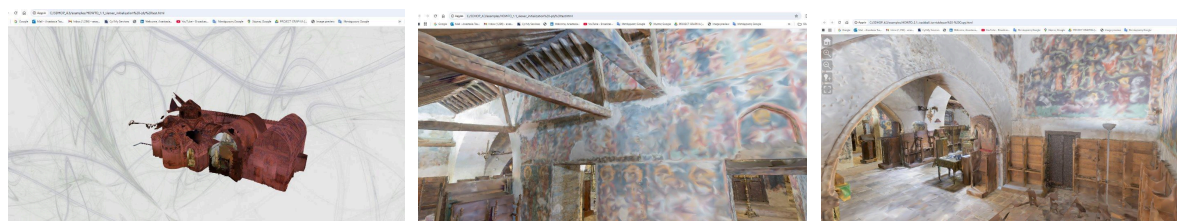


Figure 4.3: Tests in 3DHOP highlighted some issues in the textures visualisation.

The team, in collaboration with CNR tried other solutions, such as using the nexus format to create a much lighter version of the 3D model. However, this still did not deliver ideal results since the texture is not recognised and correctly displayed. Currently other trials and tests of different methods are ongoing. This issue has slightly delayed the development of the successive phase, which consists of the development of the storytelling project for the demonstrator with 3DHOP. Nevertheless, a mock-up of the storytelling project has been drafted with the selection of the tools which need to be integrated, like layers and single components visualisation, annotations, pop-up information. The pop-up will provide historical information about the development of the site and could also contain links to related studies. The team also tested the

integration of a reconstructed Historic Building Information Modelling (HBIM) model of the Monastery but the test encountered issues regarding the compatibility of Industry Foundation Classes (IFC) formats with the viewer so this approach was discounted.

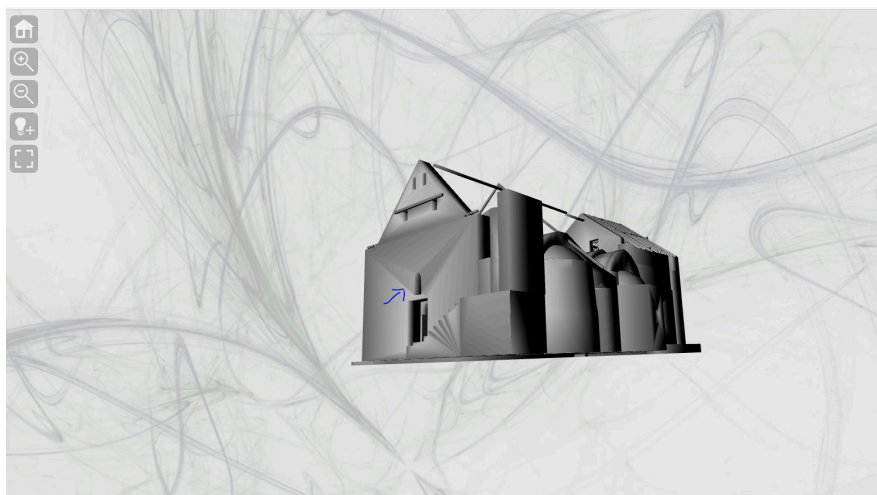


Figure 4.4: Test for the visualisation of the HBIM reconstructed model of the Monastery

Meanwhile, the embedding of 3DHOP in the EpHEMERA platform has been successfully tested for integration of the viewer and the visualisation and demonstrator. The EpHEMERA platform allows users to access and select 3D viewers according to their needs.

The demonstrator, with preliminary results of the visualisation, was presented at CAA 2025 in Athens during a session organised by ATRIUM. A paper on this case study is currently under preparation for the dissemination of our work. The development of the case study is ongoing and will be finalised during the next reporting period.

Kampanopetra Basilica, Cyprus

The great Basilica of Kampanopetra is located in the southeastern part of the ancient city of Salamis on the east coast of Cyprus. The city suffered from various natural disasters during its long history. Historical sources attribute the destruction and decline of ancient Salamis to earthquakes in the fourth century, with archaeological evidence confirming the violent collapse of buildings and the disturbed geomorphology of the site. Kampanopetra was constructed after this destruction and as part of the fourth century efforts to renew the city. The complex was affected and destroyed by the Arab raids of the seventh century, which led to the final decline of the city.

The Kampanopetra Basilica was selected as a case study representing an archaeological site and to demonstrate a 3D workflow for virtual hypothetical

reconstruction aimed at research and valorisation. In particular, the 3D reconstruction of the site¹ is driven by a series of specific objectives, including: raising awareness of the importance and value of the archaeological site at local, national and international levels; supporting learning activities, research, and studies by allowing a virtual visit and experience of the site; contributing to its digital preservation; supporting the development of qualified professionals in knowledge management and knowledge sharing regarding cultural heritage sites; and engaging hard-to-reach audiences who do not participate in cultural heritage-related activities. The choice and use of open data and software for the demonstrator was crucial. Namely, it must use dynamic 3D viewers that can be easily embedded in the EpHEMERA platform to guarantee the accessibility, sharing and collaboration needed for the preservation and valorisation of the site.

In relation to the aims of the case study, the team selected 3DHOP as the web viewer for the visualisation of the [Kampanopetra Basilica](#) (Figure 4.5). Considering that the authoring process of the historical virtual reconstruction had to reflect its interpretive limits, the development of a schematic map dividing the building's components into four categories helped to address the related challenges and hypotheses. The first category comprises surviving standing structures so that their 3D shape is detectable. The second category identifies structures which only survive in plan, meaning we can see only their footprint. Wall shape and orientation, as well as the location of columns, are extracted from the existing footprint. The third category includes architectural elements reconstructed from literature and related sketches and drawings (Roux 1998). The final category includes the elements that were created based on archaeological and architectural observations and interpretations, as no other evidence was available.

The 3DHOP implementation was designed to present all the models that were created for the Basilica as layers. The visualisation loads the Basilica and the surroundings in their current state of conservation. There is the possibility to visualise the reconstruction in transparent view and a color-coded visualisation that is based on the schematic map described above. As the color-coded model was also created in IFC format before converting to the Nexus format, it suffers from the same issues as the previous case study.

¹ The 3D virtual reconstruction of the Kampanopetra Basilica is a project of the Technical Committee on Cultural Heritage (TCCH), implemented by the United Nations Development Programme (Cyprus) in collaboration with the APAC Labs of the Cyprus Institute.



Figure 4.5: A snapshot from the interactive visualisation of the Basilica and its surroundings in 3DHOP.

In order to further improve the user interaction and allow a more engaging experience, another Web viewer ([ATON](#)²) was considered for possible future integration in the EpHEMERA platform. Hence the models were also implemented in the ATON platform (Figure 4.6). A textured model was created to provide an outline impression of how the basilica complex may have appeared in the past. The same layering was applied in ATON, but with slight alterations. For example, the extant remains were visualised in two layers, one for the basilica and one for the surrounding landscape. This allows the user to visualise better the spatial relation between the existing remains and the historical virtual reconstruction. Another key difference is the separation of the color coded model based on the schematic map. This allows the user to navigate the reconstruction based on the source used to reconstruct the various elements. In addition, the roof was moved to a separate layer giving the opportunity to examine the basilica without the roof or examine the roof elements separately. Finally, the first-person navigation mode available in ATON offers a game-like experience that can be attractive to a younger public (Figure 4.7). Tests for the integration of the ATON viewer in the EpHEMERA platform will be carried out in the near future. The demonstrator and its workflow for 3D virtual reconstructions were presented at the Digital Heritage 2025 conference and published in a peer-reviewed paper (Faka et al 2025).

² ATON is an open-source framework to create Web3D/WebXR apps interacting with CH objects and 3D scenes on the Web.

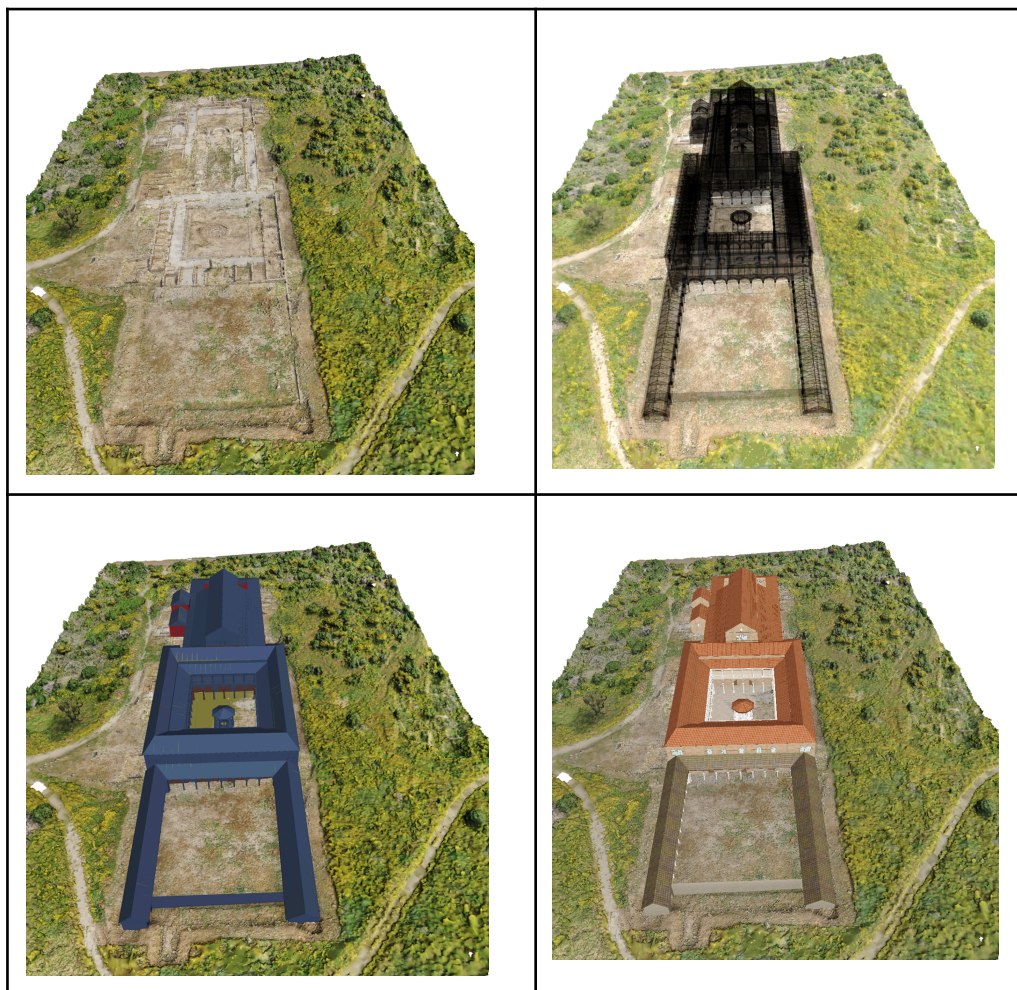
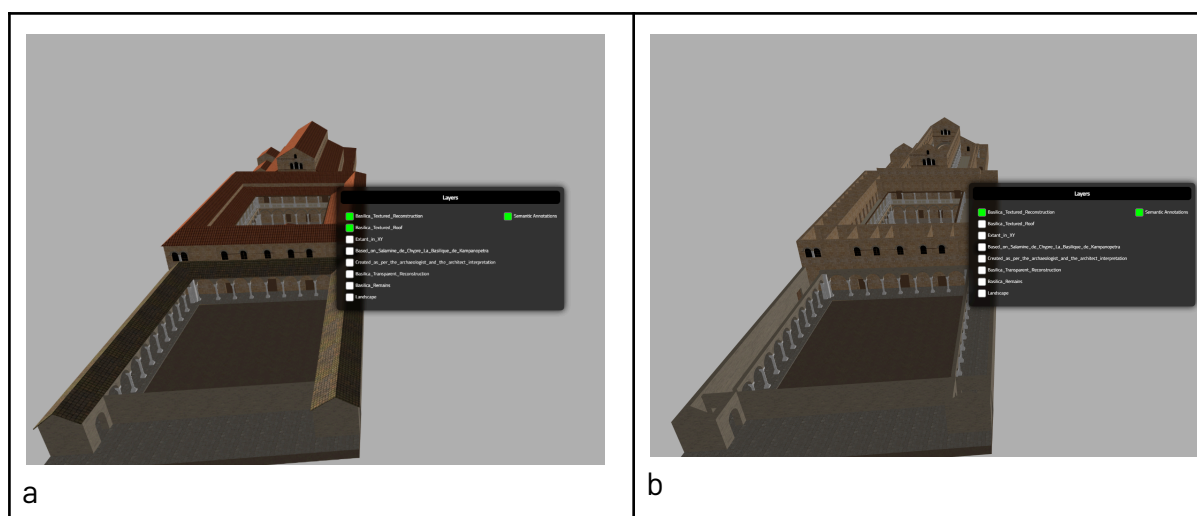


Figure 4.6: Snapshots of the different types of visualisation in ATON.



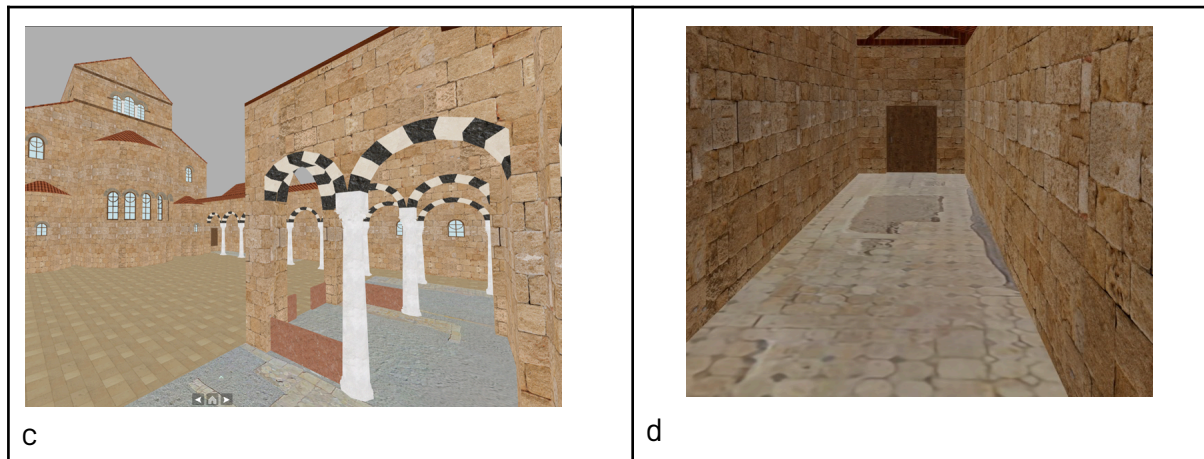


Figure 4.7. Snapshots of the virtual reconstruction visualisation: a) the reconstruction visualised with the roof and the different layers of the implementation; b) the reconstruction visualised without the roof; c) first person viewpoint; d) incorporation of the original mosaics in the textured model.

4.1.4 Viewers

Following the activities in T4.3.1, the 3DHOP viewer has been extended, introducing tools and visualisation modes. The interface of the viewer has been revamped for integration in the ARIADNE portal, adding different interface controls and options. The main focus of this task was to cope with the very diverse nature and characteristics of the indexed resources, setting up a general-purpose viewer with a lot of options. Several of these additions were required for the demonstrators, including a “see through” rendering mode that, using backface culling, makes it possible to see the inner part of 3D models, which works perfectly on buildings (see Figure 4.8). These additions to 3DHOP are now being integrated in the implementations for the demonstrators.

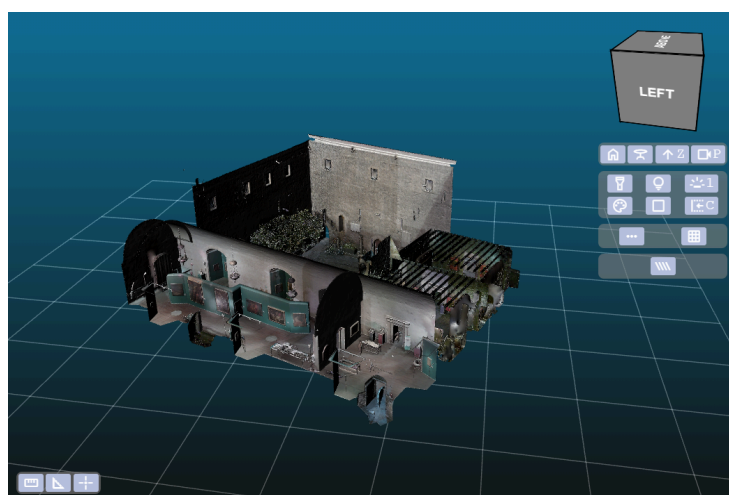


Figure 4.8. 3DHOP visualisation now includes a “see-through” mode that is especially useful for 3D models of buildings, to see the inner parts in relation to the whole structure.

4.1.5 Provision of final datasets

The demonstrators will give users the possibility to explore how 3D architectural models may function as spatial interfaces, providing access to multidisciplinary information through 3D viewers at item level. Access to the case studies will be provided via the ARIADNE portal through the EpHEMERA service, already available as an ARIADNE visualisation service, but not yet integrated in the portal. Within the T5.3 demonstrator the selected viewers will now be integrated in the ARIADNE portal through an iFrame solution, with metadata for the individual case studies uploaded to the Knowledge Base. To facilitate this, the metadata schema used in EpHEMERA has been extracted and, together with the WP3 team, an initial metadata mapping to the AO-Cat has been undertaken (Figure 4.9). In the next period, an update of the mapping with a link to the final demonstrators will be carried out.

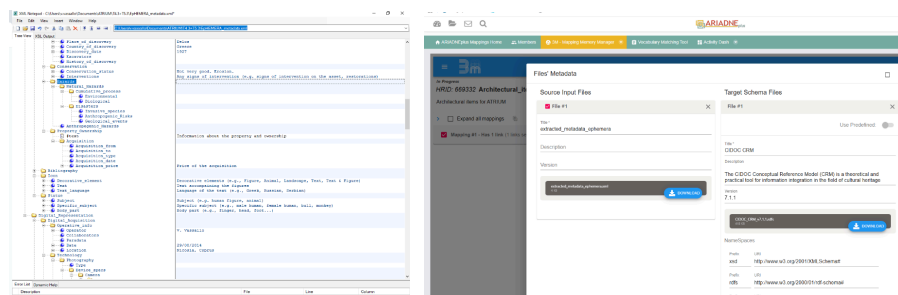


Figure 4.9: Metadata alignment for the aggregation into the ARIADNE portal and its visualisation.

4.2 Adoption of Historic Building Information Modelling (HBIM) (T5.3.2)

Summary

HBIM is a 3D information management system that has become widely used in architectural studies and building management but which is also being increasingly adopted by archaeological contractors, particularly as they have to share data with structural engineers as part of large infrastructure projects. The HBIM adoption task (T5.3.2) focuses on integrating heritage BIM datasets from the Cyprus Institute (Cyl) and Laboratório Nacional de Engenharia Civil (LNEC) to support conservation-oriented workflows, assessing their completeness across historical, geometric, and conservation data domains. Two representative case studies were developed to reflect different heritage contexts, using open IFC standards and XeoKit as the primary viewer due to its performance and interoperability. Integration with the ARIADNE portal through iFrame embedding and metadata mapping to the AO-Cat is underway to enable online visualisation and access to the final demonstrators.

Workflow

<https://marketplace.sshopencloud.eu/workflow/rnzX2P>

4.2.1 Provision of sample datasets

Both Cyl and LNEC maintain independent repositories of heritage BIM datasets. These datasets were systematically analysed and reviewed by the team with respect to their completeness and suitability for supporting the HBIM workflow. The assessment focused on three essential information domains required for HBIM implementation:

1. Historical Analysis Data: documentary, archival, and contextual historical sources;
2. Geometric Survey Data: terrestrial and aerial documentation, including laser scanning, photogrammetry, and traditional survey techniques;
3. Conservation State Analysis: condition assessments, material diagnostics, and records of previous interventions.

Considering the inherent heterogeneity of HBIM use cases, ranging from conservation planning and preventive maintenance to digital twin applications, the team adopted a general-purpose information benchmark tailored to architectural conservation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	2	HISTORICAL ANALYSIS														
3	Number	Data Name	Schedule Domain	Subject of Study	Value of Measure	Capturing Method	Raw Data Processing	Parameterization	Open BIM Framework Integration	List of compatible Vocabularies	List of compatible Ontologies	International or local standards	Comments and notes			
		Historical analysis for heritage preservation involves integrating historical data with modern digital tools and methods to understand the evolution of materials, construction techniques, and building design in relation to their context.														
3		Building Number	Label or reference assigned to a specific piece of data within a dataset	Architectural Domain: Drawing, design, etc.	The historical context provided by details of a particular building, drawing, or document is of great value for understanding the evolution of a building's form, function, and structure.	Technique or process used to capture or record data, such as drawing, photography, or scanning.	Raw data extracted from a specific source, such as a drawing, photograph, or scan.	Refers to the process of converting raw data into a format that can be used for analysis.	The Open BIM framework provides a common language for integrating data from different sources, such as architectural drawings, photographs, and digital data.	In the context of Open BIM, "Vocabularies" refer to the standardized sets of terms and definitions used to describe building information and its relationships.	In the context of Open BIM, "Ontologies" refer to the structured representations of knowledge in a specific domain, typically represented as a hierarchy of concepts and relationships.	International standards for building information modeling (BIM) and digital heritage preservation.	Add comments or notes here.			
4	A	Historical Documentation	History	Architectural records	Qualitative	Historical Data (textual)	Archival Research	PDF, DOCX, XLS, CSV	Document digitization, transcription	Historical narratives, timelines	Historical data integration, verification of building plans	CIDOC-CRM, LILO	CIDOC, CRM, EDM	(B01101, B01103)		
5	A1	Architects	History	Text - Name	Textual	Archival Research	PDF, DOCX, XLS, CSV	Historical document analysis	IPC	Uniclass, Omniclass	CIDOC, CRM, LILO	CIDOC, CRM, EDM				
6	A2	Construction Date	Timeline	Text - Description	Textual	Field Surveys, Archival Research	PDF, DOCX	Measurement, documentation	IPC	Uniclass, Omniclass	CIDOC, CRM, LILO	CIDOC, CRM, EDM				
7	A3	Movement Description	Cultural Heritage	Text - Description	Textual	Field Surveys, Archival Research	PDF, DOCX	Measurement, documentation	IPC	Uniclass, Omniclass	CIDOC, CRM, LILO	CIDOC, CRM, EDM				
8	A4	Building Materials	Material Science	Material Composition	Material Types	Textual	Sampling, Archival Research	CSV, JSON	Material classification and categorization	IPC	Uniclass, Omniclass	CIDOC, CRM, LILO	CIDOC, CRM, EDM	(B01474) (B0402)		
9	A5	Restoration History	History	Building Modification	Restoration Dates	Date	Archival Research	PDF, CSV	Restoration timeline from historical records	IPC	Uniclass, Omniclass	CIDOC, CRM, LILO	CIDOC, CRM, EDM	(B01121) (B01122) (B01123) (B01124) (B01125) (B01126) (B01127) (B01128) (B01129) (B01130) (B01131) (B01132) (B01133) (B01134) (B01135) (B01136) (B01137) (B01138) (B01139) (B01140) (B01141) (B01142) (B01143) (B01144) (B01145) (B01146) (B01147) (B01148) (B01149) (B01150) (B01151) (B01152) (B01153) (B01154) (B01155) (B01156) (B01157) (B01158) (B01159) (B01160) (B01161) (B01162) (B01163) (B01164) (B01165) (B01166) (B01167) (B01168) (B01169) (B01170) (B01171) (B01172) (B01173) (B01174) (B01175) (B01176) (B01177) (B01178) (B01179) (B01180) (B01181) (B01182) (B01183) (B01184) (B01185) (B01186) (B01187) (B01188) (B01189) (B01190) (B01191) (B01192) (B01193) (B01194) (B01195) (B01196) (B01197) (B01198) (B01199) (B01200) (B01201) (B01202) (B01203) (B01204) (B01205) (B01206) (B01207) (B01208) (B01209) (B01210) (B01211) (B01212) (B01213) (B01214) (B01215) (B01216) (B01217) (B01218) (B01219) (B01220) (B01221) (B01222) (B01223) (B01224) (B01225) (B01226) (B01227) (B01228) (B01229) (B01230) (B01231) (B01232) (B01233) (B01234) (B01235) (B01236) (B01237) (B01238) (B01239) (B01240) (B01241) (B01242) (B01243) (B01244) (B01245) (B01246) (B01247) (B01248) (B01249) (B01250) (B01251) (B01252) (B01253) (B01254) (B01255) (B01256) (B01257) (B01258) (B01259) (B01260) (B01261) (B01262) (B01263) (B01264) (B01265) (B01266) (B01267) (B01268) (B01269) (B01270) (B01271) (B01272) (B01273) (B01274) (B01275) (B01276) (B01277) (B01278) (B01279) (B01280) (B01281) (B01282) (B01283) (B01284) (B01285) (B01286) (B01287) (B01288) (B01289) (B01290) (B01291) (B01292) (B01293) (B01294) (B01295) (B01296) (B01297) (B01298) (B01299) (B01300) (B01301) (B01302) (B01303) (B01304) (B01305) (B01306) (B01307) (B01308) (B01309) (B01310) (B01311) (B01312) (B01313) (B01314) (B01315) (B01316) (B01317) (B01318) (B01319) (B01320) (B01321) (B01322) (B01323) (B01324) (B01325) (B01326) (B01327) (B01328) (B01329) (B01330) (B01331) (B01332) (B01333) (B01334) (B01335) (B01336) (B01337) (B01338) (B01339) (B01340) (B01341) (B01342) (B01343) (B01344) (B01345) (B01346) (B01347) (B01348) (B01349) (B01350) (B01351) (B01352) (B01353) (B01354) (B01355) (B01356) (B01357) (B01358) (B01359) (B01360) (B01361) (B01362) (B01363) (B01364) (B01365) (B01366) (B01367) (B01368) (B01369) (B01370) (B01371) (B01372) (B01373) (B01374) (B01375) (B01376) (B01377) (B01378) (B01379) (B01380) (B01381) (B01382) (B01383) (B01384) (B01385) (B01386) (B01387) (B01388) (B01389) (B01390) (B01391) (B01392) (B01393) (B01394) (B01395) (B01396) (B01397) (B01398) (B01399) (B01400) (B01401) (B01402) (B01403) (B01404) (B01405) (B01406) (B01407) (B01408) (B01409) (B01410) (B01411) (B01412) (B01413) (B01		

4.2.2 Case studies

- Building and Architectural Information;
- Contextual description;
- Conservation History and previous interventions.

In alignment with open standards, HBIM models are managed and exchanged in IFC format. The team conducted an initial survey of openly available IFC viewers to identify suitable tools for the project's requirements. The evaluation demonstrated that XeoKit is currently the only viable open-access option offering the necessary performance, compatibility, and extensibility. It also has a frequent update and integrates new visualisation tools. XeoKit is already integrated into Cyl's Urban Periscope platform, where it has demonstrated reliability and efficiency in rendering complex architectural datasets for the last four years.

Pilot Building Case

[This document serves as a template for collecting information on the Pilot Building, integral to the development of the HBIM workflow for the ATRIUM Project. The template includes example data to guide users in data collection and documentation.]

SECTION 1. PILOT INFORMATION					
Description of the pilot					
Provide general identity data of the building:					
Building Name	Municipal Museum of Folk Art				
Location	Limassol, Cyprus				
Address	Agioy Andreou 253				
Ownership Type	Public				
Current Owner	The Limassol Municipality				
Ownership (by chronological order)	UK military forces, Danish Canadian military detachment in Cyprus, Republic of Cyprus, Municipality of Nicosia				
Original Construction Date	Estimated 1920s				
Current Use	Museum				
Use (by chronological order)	Residential, Nursing Home, Art Gallery				
Description	The two-storey building was constructed in the early 1920s. It used to accommodate two individual residences, one in each floor. Its main façade is symmetrical and on Agioy Andreou street. The entrance to the first floor is via a stone staircase at the side of the building.				
Listing Status	Listed				
GENERAL PARCEL INFORMATION					
Zone	Plot Number	Quarter	Sheet / Plan	Parcel Area (Sq.m)	Parcel Perimeter (m)
Πα6	1075	AGIA TRIAS	54/580601	508	94

Figure 4.11: Pilot Building Case

To support the project's technical workflow and the upcoming demonstrators, ongoing integration tests are being performed using an iFrame-based embedding solution, allowing the team to evaluate interoperability, responsiveness, and user interaction within different platform environments. Currently the Cyl team is working on isolating the viewer or creating an embedded link for clarity purposes.

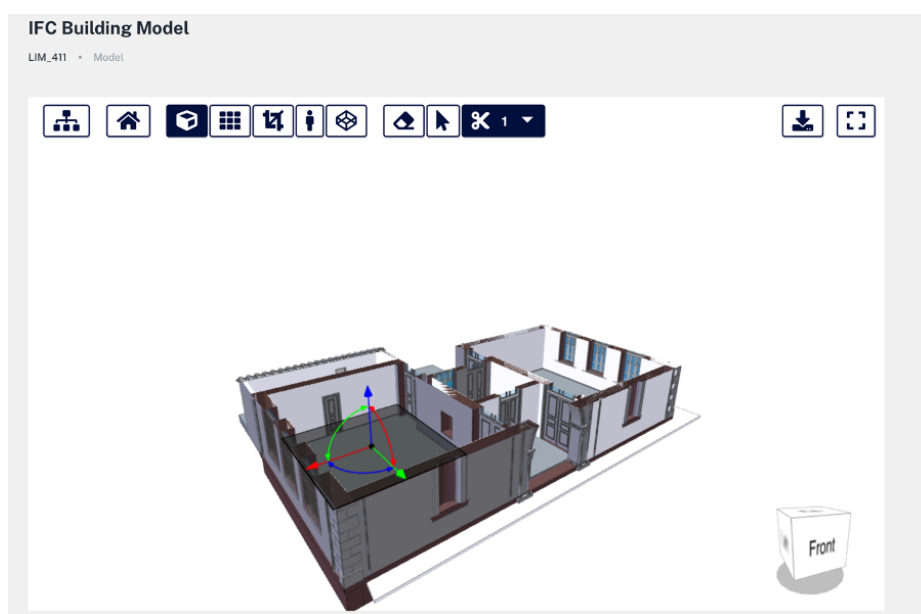


Figure 4.12: In Urban Periscope Portal

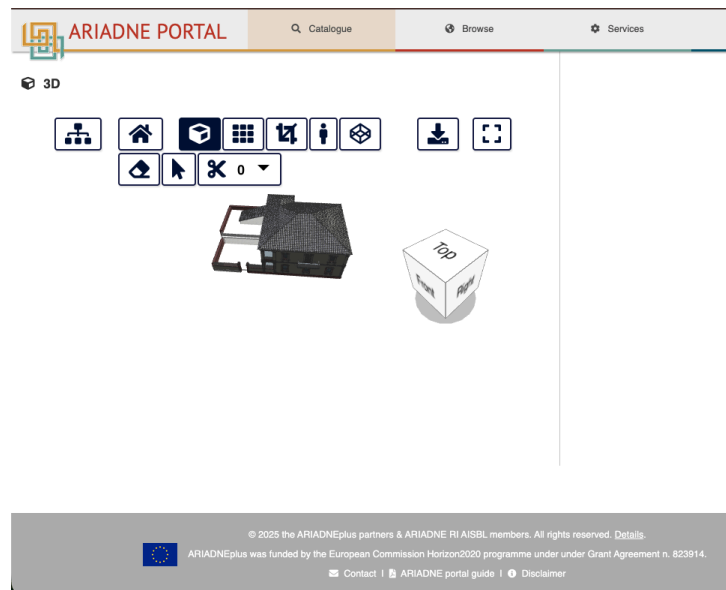


Figure 4.13: In the ARIADNE Portal

4.2.4 Provision of final datasets


Access to the individual HBIM case studies selected by Cyl and LNEC will be provided via the ARIADNE portal by means of the related services and an iFrame solution. In this case, too, the metadata must be aggregated in the Knowledge Base. The metadata schema currently used by the Cyl Urban Periscope platform has been extracted, and an update is ongoing to guarantee the presence of mandatory fields for mapping to the AO-Cat. In the next period, the metadata related to the LNEC case studies will be extracted, along with the metadata schema mapping and the link to the final demonstrators for both institutions, to allow them to be visualised in the ARIADNE Portal. The demonstrator will provide examples of how researchers or heritage managers will be able to search for specific cultural heritage assets, including standing buildings as well as reconstructions, read descriptive information, and then use the viewers to visualise and interrogate the models.

5. T5.4 Sound-Based Demonstrators

Introduction

Field archaeologists traditionally use paper context sheets for documenting their findings, but these can be time-consuming to prepare and may pose practical difficulties in a wet and muddy field environment. With the growing sophistication of speech-to-text technology we therefore decided to add an additional data type to the ATRIUM workflows: voice recordings. This was facilitated by the participation of CLARIN partners in ATRIUM who provided technical expertise. The main objective of this task in WP4/5 is therefore to assist data entry using automatic speech recognition (ASR) technologies, enabling archaeologists to document context sheets via speech instead of handwriting, improving efficiency and accessibility. T5.4 focuses on developing and testing a sound-based demonstrator to assist archaeologists in the field, addressing challenges they face in recording context data. The outputs from this task can then be fed into the workflow developed in T4.1 for texts and showcased via transcribed sounds recordings in the ADS Data catalogue and ARIADNE portal. This task connects with the 'text from speech' section of Task [5.1](#).

The workflow underlying this demonstrator is available here:

 [T4.4 - Sound-based workflow](#)

To achieve the objectives, the team at ATHENA-RC and UoY-ADS developed a web application that allows users (field archaeologists) to create context sheet forms and enter data, such as context sheet number, description, and interpretation, i.e., free text fields that are typically the most time-consuming. By using speech, archaeologists can potentially provide more detailed and natural responses without the constraints of handwriting.

In summer 2025, the first field trials were conducted at the Toumba Serron excavation ditch in Greece. However, during the initial tests, the whisper-based ASR model that was used proved to have unacceptable latency, making it unsuitable for real-time field use. To address this, the team switched to a low-latency model developed by Kyutai, which has proven to be much faster and more effective in internal tests. Recently, the team finalised changes to the user interface to improve usability, and the web application was released, with the following [source code](#) and [demo deployment](#).

5.1 Provision of sample datasets

Voice notes had not previously been used in the field by UK archaeologists, so a "Context Sheet Reading Party" was organised at the University of York. During this event, eight volunteers read aloud prompts based on existing archaeological context sheet descriptions, which were then used to create sample datasets for the ASR model (total duration: 2 hours, 23 minutes, 31.51 seconds). For this task, a voice recording web app was deployed (see the figure below), which allowed participants to record their readings through their browser.

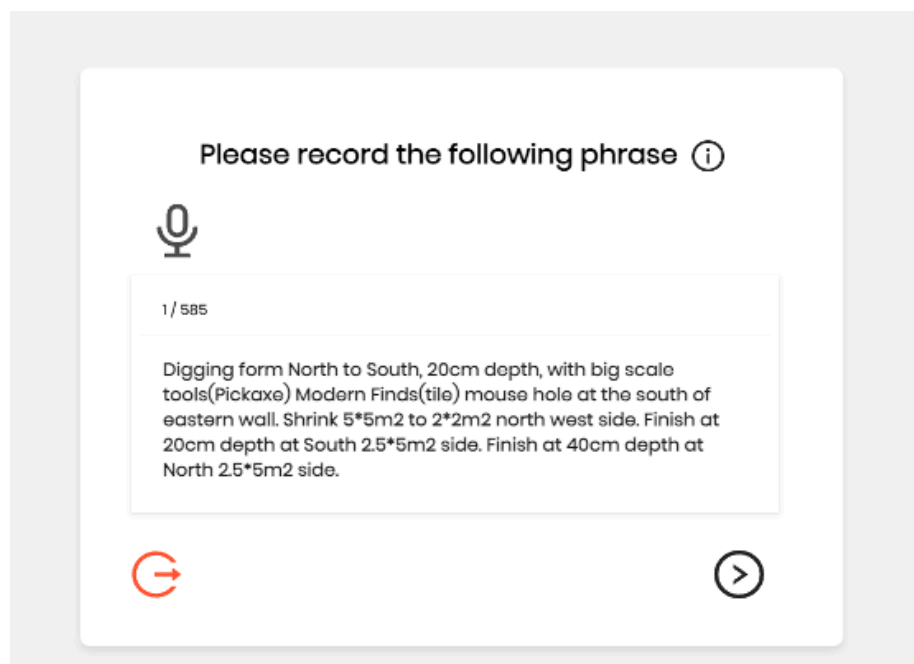


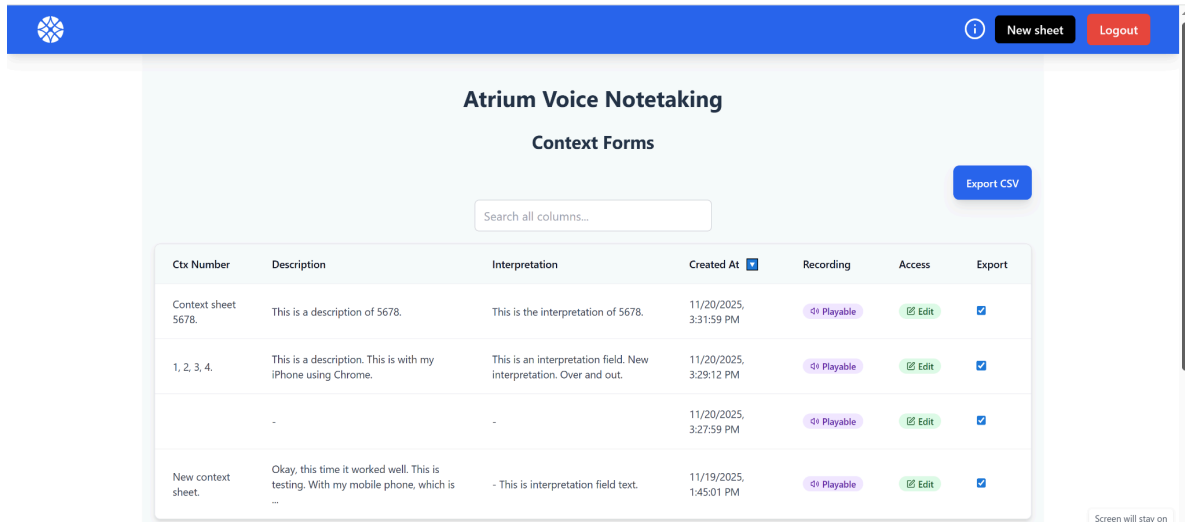
Figure 5.1: Voice recording web app interface

5.2 The Web Application

A web application was developed that allows archaeologists to create and fill out context sheets using speech. This tool is designed to address common challenges, such as writing under harsh field conditions (e.g., dirt, difficult weather conditions) or accessibility issues (e.g., dyslexia, tendonitis). The application allows users to record the context sheet number, description, and interpretation, through speech, potentially allowing them to be more creative and less concise compared to interrupting their work and providing handwritten entries.

The web application is available in two ways: as a [demo deployment](#) (see dashboard and interface for adding new context sheets below), and as an open-source project hosted

on [GitLab](#), where anyone can clone and deploy it on their own servers. Additionally, the application includes an export feature that generates context sheet data in CSV format, which can be ingested and archived and disseminated by providers such as ADS or AIR (Archaeological Interactive Reports system).



Atrium Voice Notetaking

Context Forms

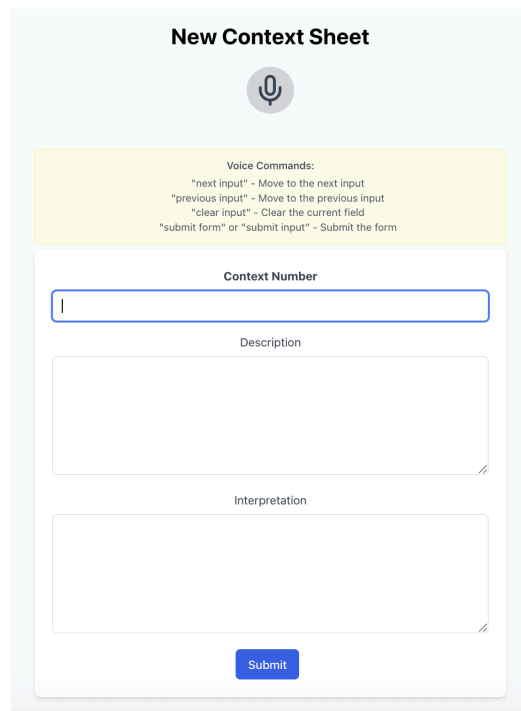
Search all columns...

Export CSV


Ctx Number	Description	Interpretation	Created At	Recording	Access	Export
Context sheet 5678.	This is a description of 5678.	This is the interpretation of 5678.	11/20/2025, 3:31:59 PM	Playable	Edit	<input checked="" type="checkbox"/>
1, 2, 3, 4.	This is a description. This is with my iPhone using Chrome.	This is an interpretation field. New interpretation. Over and out.	11/20/2025, 3:29:12 PM	Playable	Edit	<input checked="" type="checkbox"/>
-	-	-	11/20/2025, 3:27:59 PM	Playable	Edit	<input checked="" type="checkbox"/>
New context sheet.	Okay, this time it worked well. This is testing. With my mobile phone, which is ...	- This is interpretation field text.	11/19/2025, 1:45:01 PM	Playable	Edit	<input checked="" type="checkbox"/>

Screen will stay on

Figure 5.2: User dashboard listing completed context forms, with options to edit, delete, and export



New Context Sheet



Voice Commands:

- "next input" - Move to the next input
- "previous input" - Move to the previous input
- "clear input" - Clear the current field
- "submit form" or "submit input" - Submit the form

Context Number

Description

Interpretation

Submit

Figure 5.3: Interface for adding a new context sheet using voice commands.

5.3 Provision of final datasets

As part of the Demonstrator, the voice transcriptions produced from the task will be archived and made available as part of the digital archive of the field project. These transcriptions will be held by ADS, supporting further research, reproducibility, and future improvements to archaeological speech-recognition tools. The data will then be used in a further task in T4.1 to tag the text with enhanced metadata according to standard controlled vocabularies. In turn, the metadata will be aggregated to the ARIADNE Knowledge Base, so that the recordings and transcriptions are discoverable through the portal as sound data.

5.4 Next steps

Workflow:

- Finish the workflow description and publish it in the SSHOC Marketplace.

Demonstrator:

- May 2026: Field data capture at the Skipsea fieldschool.
- Improvements to app and model based on feedback from the fieldschool.
- Archive the produced dataset at ADS.
- Disseminate the context sheets and their enriched metadata through the ARIADNE portal.

6. T5.5 Geospatial Demonstrators

Introduction

Archaeology is unusual for the reliance it places on geospatial location of its primary data. This task uses “place” as a central element to link archaeological and cultural heritage datasets with datasets from other disciplines. It demonstrates collaborative map annotation through Recogito Studio using Rigas Velestinlis’ Charta of Greece and related historical maps, showcasing semantic linking and multi-repository annotation workflows. A second demonstrator integrates spatial data from Swedish and British case studies within the ARIADNE portal, culminating in an interactive Peripleo map that visualises interdisciplinary connections across time and space.

A dedicated instance of Recogito Studio has been deployed and extensively tested for collaborative map annotation workflows. Sample datasets from Polish and Austrian repositories have been identified and enriched with metadata, with the Charta of Greece (1797) selected as the focal demonstration case. Spatial datasets from Swedish (Database Sweden 1570-1810) and British (St. James burial ground) repositories have been prepared for integration into the ARIADNE portal.

6.1 Collaborative Map Annotation (T5.5.1)

6.1.1 Overview

Task 5.5.1 aims to demonstrate the [collaborative map annotation workflow](#) developed in Task 4.5.1 by applying it to selected historical maps from multiple repositories containing rich, multi level information with the [Charta of Greece](#) (1797) by Rigas Velestinlis as the focal point. The demonstrator primarily uses [Recogito Studio](#) to showcase collaborative practices, with particular emphasis on the semantic linking of place names to authoritative gazetteers.

6.1.2 Work achieved in the task

The task has so far focused on defining the research case for the demonstrator (centered on the Charta of Greece), identifying suitable datasets, evaluating and selecting annotation tools, and setting up a dedicated instance of Recogito Studio. The original plan was to use the tool Recogito. However, due to developments the tool is no longer supported. As such, the team switched to Recogito Studio, which required extensive testing as the tool is still in active development.

During the preparatory phase of the demonstrator, the team reviewed sample datasets from partners, enriched them with additional metadata, and identified maps relevant to the Charta. Recogito Studio has been tested thoroughly to ensure it meets the needs of collaborative map annotation, and sample data has already been annotated to validate the workflow.

An initial presentation of the workflow and demonstrator concept was also delivered at the CAA 2025 conference in Athens:

- [07 CAA2025_ATRIUM_MapAnnotationWorkflow.pptx](#) (version archived on [Zenodo](#))

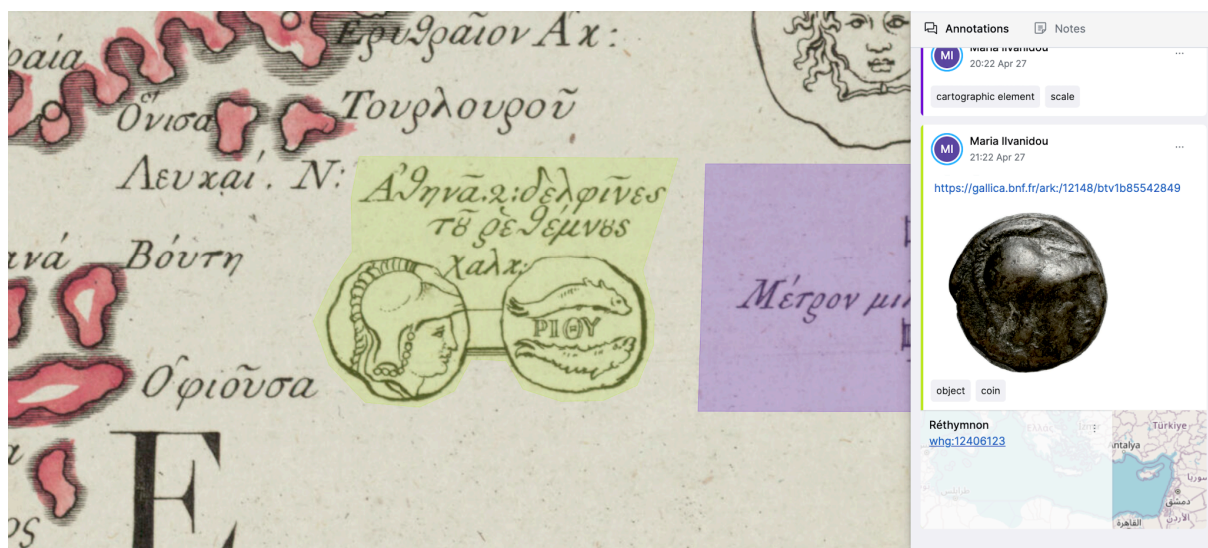


Figure 6.1: Example of annotation of the Charta of Greece in Recogito Studio

6.1.3 Timeline


- Finalise the workflow developed in Task 4.5.1, along with any complementary workflows, and integrate the relevant elements into the demonstrator.
- Further refine the selection of maps from the enriched data samples, align them with the chosen research scenarios, and define the final dataset.
- Select the specific annotation guidelines/criteria and examples to apply to the Charta and the other maps in the demonstrator.
- Determine the final structure and presentation of the demonstrator.
- Carry out all annotations needed to showcase the selected research scenarios.
- Prepare the training materials and a training session.
- Prepare the demonstrator documentation to support WP5 activities and develop the dissemination material.
- Archive the demonstrator in a publicly available repository to ensure its long-term preservation, accessibility, and broad dissemination.

6.1.4 Provision of sample datasets

Regarding the dataset, at first datasets provided to ATRIUM by ADS and ARUP/ARUB were reviewed to determine whether they could be used for subtasks T4.5.1/5.5.1 and T5.5.2. This review resulted in a short report that concluded that none of these datasets were suitable for map annotation:

- [A review of data samples for the ATRIUM project – Internal Report](#)

In parallel, we prepared a [questionnaire](#) and circulated it to ATRIUM partners to collect any historical map datasets that could support the demonstrator. Since the demonstrator needs to support annotating maps from multiple repositories, we needed examples from more than one source. Two sample datasets were provided by partners:

-  `sampleMapData_PSNCR.xlsx` – a dataset of roughly 2000 public-domain maps from one of the data providers of Federacja Bibliotek Cyfrowych (FBC, the Polish aggregator run by PCSS), specifically the Digital Library of the University of Wrocław.
- [Woldan Collection \(OEAW\)](#) – a dataset of about 1200 historical maps from the Woldan Collection at the Austrian Academy of Sciences.

6.1.5 Choice of tools – Recogito Studio

Since Recogito, the tool of choice at the proposal stage, is now outdated and no longer supported, we assessed alternative annotation tools against the key requirements established in T4.5.1 (collaborative features, IIIF compatibility, integration with gazetteers, etc). In total we examined around 40 annotation tools against 12 criteria and concluded that Recogito Studio, the successor to Recogito, best meets these needs and is the most suitable option for collaborative reusable map annotation. The results of this evaluation are documented in:

- [Assessment of Annotation Tools for T5.5.1](#) (v.2)
- [ATRIUM T5.5.1 Annotation tools overview](#) (v.2)

Therefore, the primary tool for the demonstrator is [Recogito Studio](#), supported by an extensible plugin system, which allows for the annotation of diverse types of materials and geolocation of place names in a collaborative environment. Annotations can be exported in interoperable formats and models, such as the [Web Annotation Data Model](#). Recogito Studio will be complemented by additional tools for specific tasks. For example, for dealing with vocabularies, the [Vocabs](#) services maintained by the Austrian Academy of Sciences will be used, which allow for the collaborative editing and publication of SKOS-based controlled vocabularies.

Recogito Studio is open source software that can be deployed on your own machine. One instance of Recogito Studio is provided by the Pelagios Network, a community of

people using Linked Data to investigate the past (<https://pelagios.recogito.studio>). This is where the first systematic annotation tests were performed. In addition, ATRIUM decided to deploy their own instance, hosted by the Austrian Academy of Sciences, to have more flexibility with regard to installed plug-ins and allocation of resources. Deploying the software on the Austrian Academy of Sciences (OEAW) servers also contributed to the identification of potential issues in the installation process and served as a test case for the developers of Recogito Studio.

As Recogito Studio was newly released and still under active development, we tested its functionality extensively in real annotation scenarios. Testing took place after each major update and led to specific reports. Any issues identified were shared with the Recogito Studio team and have since been resolved. The results of the testing are documented in:

- [Recogito feedback for devs](#)
- [ATRIUM - Testing the Updated ATRIUM Recogito Studio Instance](#)

The feedback provided to the developers as well as the active financial support by the ATRIUM partners also led to the improvement of specific aspects of the graphical user interface, the development of new functionality, and the introduction of closer integration with other software components (such as Named Entity Recognition services).

6.1.6 Case study

As it was important for the collaborative map annotation workflow to support both simple and complex scenarios, we designed the demonstrator around a strong and compelling research case and set out to identify and prepare suitable historical maps for annotation.

We selected the **Charta of Greece** by Rigas Velestinlis as the main focus, a landmark 18th-century map renowned for its dense historical, archaeological, and cultural content. Engraved in Vienna and published in 12 sheets in 1796-1797, the Charta covers a broad area from the Danube to Crete and contains more than 5800 place names, multiple historical layers, city plans, coins, mythological elements, and detailed commentary. Its complexity makes it an ideal testbed for exploring the challenges of historical map annotation and serves as a representative example of deep mapping. Working with the Charta helps clarify and demonstrate practical issues such as defining annotation goals, selecting place types, integrating suitable gazetteers, handling ancient and modern toponyms, annotating non-place features, defining vocabulary needs, and managing ambiguities.

In parallel, examining the Charta alongside other historical maps of the same regions (both earlier and contemporary), selected from the sample datasets, provides horizontal insights relevant to annotation and geolocation across diverse repositories. This includes, for example, identifying variations of the same place across periods and map types or tracing possible sources that informed the creation of the Charta.

All background material and preparatory notes on the Charta of Greece have been collected in a dedicated document:

- [Rigas' The Charta of Greece](#)

6.1.7 Provision of final datasets

The demonstrator will be built around Rigas' Charta of Greece, using the digitised and stitched version provided by the Harvard Library:

- <https://iiif.lib.harvard.edu/manifests/view/ids:24146534>

In addition, the Charta of Greece will be compared and jointly annotated with earlier and contemporary historical maps of the same regions drawn from the sample datasets. To support this, we enriched the sample datasets with information on geographic coverage, date, links to IIIF manifests or specific pages on atlases, and their relevance to the Charta. From the roughly 2000 maps in the PCSS sample, around 450 were identified as relevant to our case study. From the nearly 1200 maps in the OEAW sample, about 50 were identified as suitable for the demonstrator. Further refinement will be carried out to select the most appropriate examples. All related information is recorded in the following documents:

- [PSNC-Maps_Charta-Relevance](#)
- [OEAW-Maps-pre1800_Charta-Relevance](#)

We have also started an informal collaboration with Eleni Gadolou (Digital Asset Manager at the British School in Athens): first, for developing an annotation [vocabulary for historical maps](#); second, for providing [further context](#) of the work on gazetteer creation (and its possible complementarity with the map of Rigas).

6.2 Using place to connect multiple disciplines across the Arts and Humanities and beyond (T5.5.2)

6.2.1 Overview

The goal of this task is to demonstrate the value of the integration of datasets from multiple research domains such as the humanities, environmental studies, and archaeology, using place, or spatial data, within the ARIADNE portal. This demonstrator

also links to [work in WP2](#) on enhancing the visibility, accessibility, and engagement with stakeholders normally outside the usual audience of the ARIADNE portal by providing a better interface, using historical and archaeological datasets with spatial data from UoY-ADS and SND and the [workflow defined in T4.5.4](#).

ADS and SND have provided the datasets used in the demonstrator, including the “Database Sweden 1570–1810” combined with historical parish polygons from the Swedish National Land Survey, and two linked historical and archaeological datasets relating to St James’ Burial Ground in London, comprising both the HS2 excavation archive and citizen-science transcribed burial records. For the Swedish case study, Östergötland parish data were merged into broader agrarian categories, linked to parish geometries, and combined with a filtered selection of archaeological and heritage datasets from SND and Riksantikvarieämbetet, while for St James’ Burial Ground, residential address fields from the burial registers were geocoded to create geo-referenced point records that can be mapped and connected to excavated burial contexts.

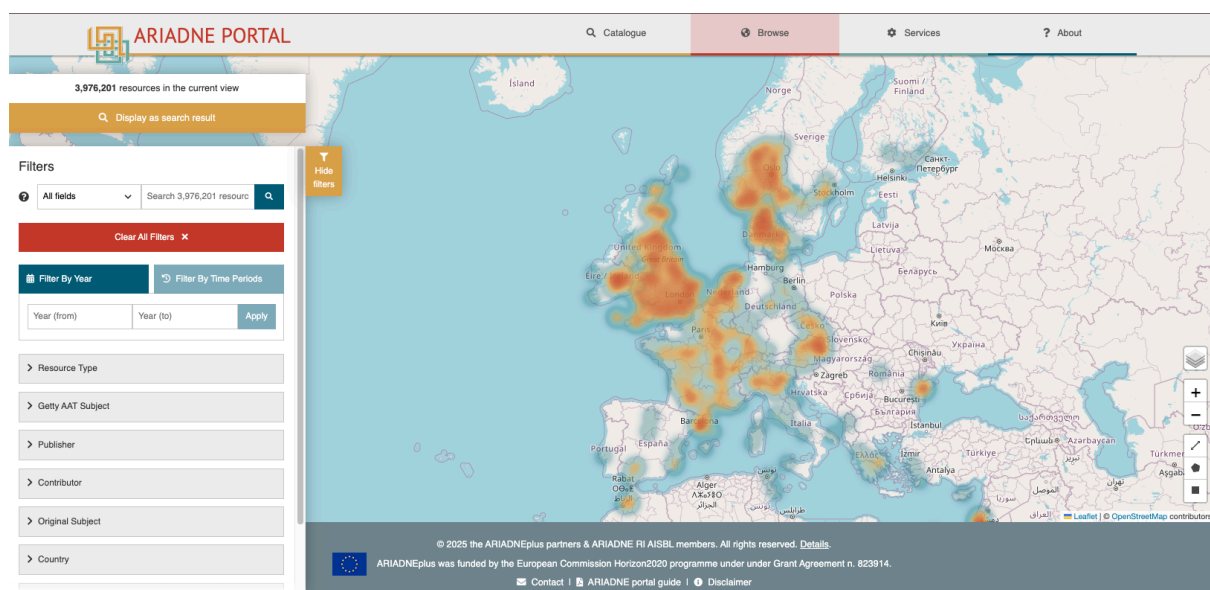


Figure 6.2: Resource distribution in the ARIADNE Portal.

6.2.2 Provision of datasets

At the start of the project, ADS and SND provided the datasets that will be used in the demonstrators. SND provided the main Swedish dataset, *the database Sweden 1570–1810*, which will be used for their case study. It was processed, archived, and published in September 2013, and can be found on the [Swedish National Database](#). This database was combined with a geopackage containing polygons over historical parishes from the [Swedish National Land Survey](#) (“Socken och stad Nedladdning,

vektor"). For the ADS case study, the team is using two different datasets, one historical and one archaeological that relate to St.James burial ground in London. The archaeological dataset is the outcome of major excavations carried out in advance of the new HighSpeed2 rail terminal at Euston. It was processed, archived, and published in August 2025 and is available on the [ADS archive](#). The linked historical dataset comprises spreadsheets created from volunteer 'Citizen Science' [transcription of burial records](#) of the burial ground between 2021 and 2023.

6.2.3 Case studies

Database Sweden 1570-1810

For the Swedish demonstrator, the case study focuses on agraro-historical data from between 1570 and 1810 in the county of Östergötland. The database relates to nationally recorded data digitised from church books and historic tax assessment lists, which include parish-level data on agricultural statistics, among others. The Östergötland sample itself contains 153 historical parishes, but it allows for a relatively contained area of study. It also overlaps with other archaeological datasets housed at SND, which the team intends to combine with, as well as datasets from the Swedish National Archive, amongst others. Historical parish names from the Östergötland data were cross-referenced with a geopackage of shapefiles, provided by [Lantmäteriet](#) (the Swedish National Land Survey), giving the outline and area of each parish according to their historical boundaries. For this demonstrator some of the variables in the Östergötland data were merged into larger categories, and then attributed to the respective parish as metadata. The agraro-historical categories that were used at the time of recording were deemed to be unnecessary for the purpose of the demonstrator and would require translating. Merging of several categories offered a viable solution e.g. 'oxar' (oxen), 'stut' (steers), 'kor' (cows), 'kvig' (heifer) and 'kalvar' (calves) were merged under the category 'cattle' for the sake of simplicity.

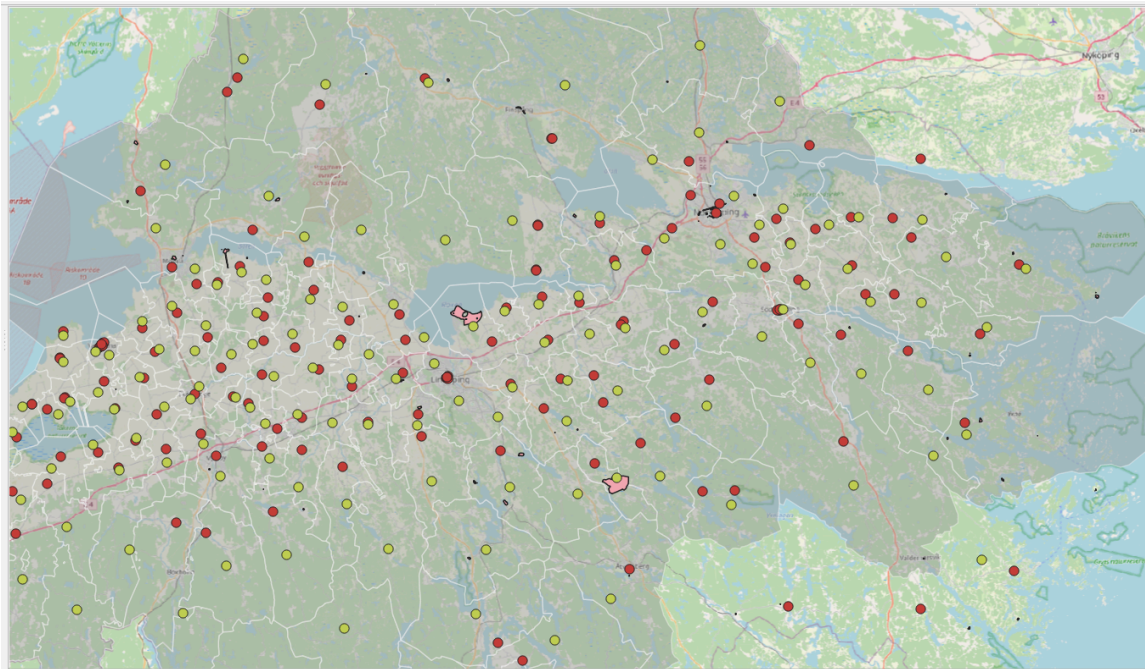
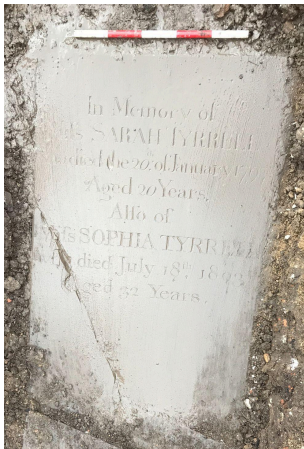


Figure 6.3: Östergötland pilot data plotted in the GIS interface.

One of the reasons that Östergötland was selected for the case study is because of the availability of combinable data within the SND catalogue. Selecting Östergötland and date range between 16th century and 19th century as a filter term in SND's catalogue on researchdata.se returned 410 usable archaeological datasets. Of these 410 data entries, 25 were selected for use in the demonstrator. This was based on the availability of digitised versions of the relevant parish church books and tax records through Riksarkivet (The Swedish National Archive). The amalgamated shapefiles will be combined and then plotted as a layer within the geopackage. Further to this, two datasets from Riksantikvarieämbetet (the Swedish National Heritage Board) were filtered and added to the geopackage to strengthen the interdisciplinary element of the demonstrator. One dataset concerned point data for historic buildings, and the other concerns polygons of conservation areas within the boundaries of the historic parishes.

Related datasets were also explored. When investigating the plausibility of combining the demonstrator with environmental data, it was found that there was only one SEAD entry applicable to the Östergötland case study, which was deemed insufficient. In contrast, a shapefile of open access archaeological excavation data, made available from Riksantikvarieämbetet provided too much data for the purposes of the demonstrator and as such it was decided to focus on the archaeological datasets available through SND.

St. James' Burial Grounds, London



The companion UK case study provided by the ADS focuses on the St James' Burial Ground, in Euston, London. Archaeological excavations were undertaken in advance of the construction of the new Euston rail terminal (High Speed Two Ltd., MOLA Northampton 2023). ADS is currently working on organising the archaeological dataset for dissemination through a special interface. When this is ready, the team can export the structured data and import to the ARIADNE Portal for the demonstrator.

Figure 6.4: MOLA Headland Infrastructure, (2018) Area C, Context: 101035; Structure [digital object]. York: Archaeology Data Service [distributor]. [Object ID: 2819236](#)

The historical dataset comprises cemetery church records of the same plots as those excavated. The registry was scanned and transcribed as part of a [Zooniverse citizen-science project](#) external to ATRIUM, then deposited at ADS as structured spreadsheets, which include burial dates and residential addresses for many individuals. As these address fields can be matched to contemporary or historical street locations using standard geocoding and GIS techniques, the team created geo-referenced point records for all individuals whose place of residence can be reliably located, enabling the historical data to be plotted on a map and connected with the excavated burial contexts.

6.2.4 Next steps

Over the remaining two years (M25–M45), the team will:

- Finalise the T4.5.4 workflow and publish it on the SSHOC Marketplace.
- Fully incorporate linked Riksarkivet data and SND datasets within the Östergötland geopackage.
- Hold an in-person workshop at ADS early 2026 to finalise presentation of datasets.
- Prepare a special collection interface for the St-James Burial Grounds archaeological dataset.
- Publish records of both case studies in the ARIADNE portal to increase their findability.
- Archive the new datasets at SND and ADS.
- Prepare a Peripleo map for the SND dataset.

6.2.5 Peripleo map

In order to maximise value from the work and to enrich this demonstrator with other dissemination means, the team has decided to publish the resulting map for the “The database Sweden 1570–1810” case study using Peripleo so users can better see the result of combining layers from multiple sources into a research map. *Peripleo* is a lightweight open-source tool for the mapping of things related to place. Originally developed by the Pelagios project (2011–2019), it was reconceived as a browser-based tool in 2022 as part of the British Library's [Locating a National Collection](#) project (LaNC) for the discovery and spatial visualisation of collection data. Various instances of it may now be found, including a DIY tutorial on github. Critically, Peripleo maps are easy to produce and disseminate as they use the adapted and simple [Linked Places Format \(LPF\)](#) or GeoJSON format for the dataset. Additionally, hosting a Peripleo map is simple and sustainable since the website created is static.

7. ARIADNE Ontology (AO-Cat) updates

7.1 Introduction

[The Ao-Cat ontology](#) is used within the ARIADNE infrastructure to harmonise and map different heritage metadata schemas to aggregate datasets from collaborating providers and improve their findability through the ARIADNE portal.

In ATRIUM, the ontology has been extended to better support the five target data types (text, images, 3D, sound, and geospatial) through new properties for declaring each resource's data type and data format, using controlled vocabularies to increase the findability of the associated datasets. Further changes, such as the introduction of AO_Digital_Media and an explicit AO_Service class, ensure that not only data but also visual materials and services are modelled to foster discoverability.

7.2 Changes to the Ontology

7.2.1 Data Types

In order to support the data types used in the ARIADNE workflows and demonstrators in the ARIADNE portal, some changes had to be introduced in the ontology. Two properties have been introduced in order to associate an ARIADNE data resource with a type and a format:

- The property `has_data_type` has as domain an `AO_Individual_Data_Resource` and range an `AO_Concept`. It associates an ARIADNE data resource with a term from a controlled vocabulary that specifies the data type of the resource. The data type refers to the categorisation of data based on its form, such as text, image, audio, video, etc. The terms used to describe the data type are defined in a controlled vocabulary of resource types specific to ARIADNE (e.g., "3D", "Image", "Audio", etc.).
- The property `has_data_format` has as domain an `AO_Individual_Data_Resource` and range an `AO_Concept`. It associates an ARIADNE data resource with a term from a controlled vocabulary that specifies the data format of the resource. The data format refers to the specific file format or encoding format used to store or represent the data, such as CSV, JSON, XML, JPEG, MP3, etc. The term will be selected from the MIME type, a standard vocabulary of formats providing a shared way of identifying the format of a resource on the internet. For example, the MIME type for a CSV file is "text/csv", and the MIME type for a JPEG file is

“image/jpeg”. The MIME type can also be used to help determine how to process or display the data resource.

7.2.2 Visual Resources

AO-Cat supported images as visual items up to version 1.2.1. In version 1.2.2 of AO-Cat Digital Media were introduced since additional requirements from ATRIUM have emerged regarding the structured representation of visual items within the ontology. It concerns the possibility of formally associating a resource with one or more digital media elements that visually describe it. Among these, a specific indication should be provided for cases where a single visual component is preferred over others, ensuring consistent representation across different parts of the Portal. A distinction is thus made between a general visual component and a primary visual component.

Furthermore, a differentiation is needed between visual media and visual images. Visual media encompass a broad range of digital assets that provide a visual representation of a resource, including images, videos, 3D and other media formats. Within this category, visual images specifically refer to static digital images used to illustrate a resource. This distinction ensures that different types of media can be accommodated while maintaining clarity in cases where only static images are relevant, such as for thumbnail previews.

To maintain a reference to the original source of a visual item while allowing its integration into the system, a mechanism should be in place to link any copied media to its counterpart on the provider’s site. This ensures traceability and supports a flexible approach where media can either be stored within the system or retrieved dynamically at display time. These requirements guarantee that visual components are consistently managed and aligned with best practices for interoperability, facilitating the integration of digital media while maintaining clear provenance and adaptability in storage and retrieval strategies.

The class `AO_Digital_Media` has been introduced in AO-Cat to model various types of digital content, including images, videos, and other media used for illustrating resources in the AC. `AO_Digital_Media` is a direct sub-class of `AO_Individual_Data_Resource` and of `crm:PE18 Dataset`, ensuring alignment with existing semantic structures.

The property `is_mirror_of` is used to associate a digital media instance in the AC with its corresponding version on the provider’s site. The domain of this property is `AO_Digital_Media`, while its range is `rdfs:Resource`. Although this property is not mandatory, it is functional, meaning that each digital media instance can be linked to at most one original source. To establish relationships between resources and digital

media, the property `has_visual_component` is defined, with `AO_Resource` as its domain and `AO_Digital_Media` as its range. This property associates a resource with any type of digital media that visually represents it. It is not mandatory, and no cardinality constraint is defined, allowing a resource to have zero, one, or multiple visual components.

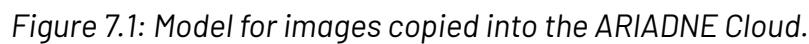
In cases where a single media instance needs to be prioritised, the property `has_primary_visual_component` is introduced as a sub-property of `has_visual_component`. It shares the same domain and range but is functional, ensuring that a resource can have at most one primary visual component. This allows providers to specify a preferred media instance for consistent representation across the AC.

Concerning the more specific case of digital images, the `AO_Digital_Image` class has the purpose of modelling thumbnails and other digital images in the AC. `AO_Digital_Image` is a direct sub-class of `AO_Digital_Media`.

Property `is_mirror_of` is used to associate a digital image in the AC with the image it is a copy of on the provider's site. Property `has_visual_component`, having `AO_Resource` as domain and `AO_Digital_Media` as range, can be used to also associate a resource with a digital image that is part of the resource and that provides some visual information about the resource. In order to let providers select a primary image to be associated with the resource, the `has_primary_visual_component` property can be used being functional (i.e., a record can have at most one primary image).

The diagram below illustrates the model for images that are copied into the AC, using a coin in the [DIME database](#) as an example. In the diagram:

- Classes are in light blue, while arcs are labelled by the property corresponding to their colors, demonstrated in the bottom part of the figure.
- `ai:idr_106900` is an individual data resource, namely a record provided by the project DIME for sharing through the ARIADNE infrastructure. In the AC, this resource:
 - refers to the coin `ai:coin_106900`, an instance of `AO_Object`
 - has the digital images `ai:di_1` to `ai:di_4` as visual components; these images are the same as the DIME resources `dime_1` to `dime_4`, respectively.
- The actual images are obtained via dereferentiation of the corresponding URIs by the appropriate web server, either on the ARIADNE infrastructure (for the URIs of the form `ai:di_n`) or on the DIME project site (for the URIs of the form `dime_n`).



For generality, the model does not make any assumption about the relationships between the images that are part of the record (ai:di_1 to ai:di_4 in the example) and the object (ai:coin_106900) that the record refers to. If desired, such relationships can be asserted as additional information. In the above example, in CRM terms any of the four images that are the content of the digital images ai:di_j is a visual representation of the coin; this can be asserted by using property P138 represents between the coin and the content of each digital image. Such level of detail is not necessary to capture the requirements stated above, therefore the classes and properties for representing this information are not part of the present model.

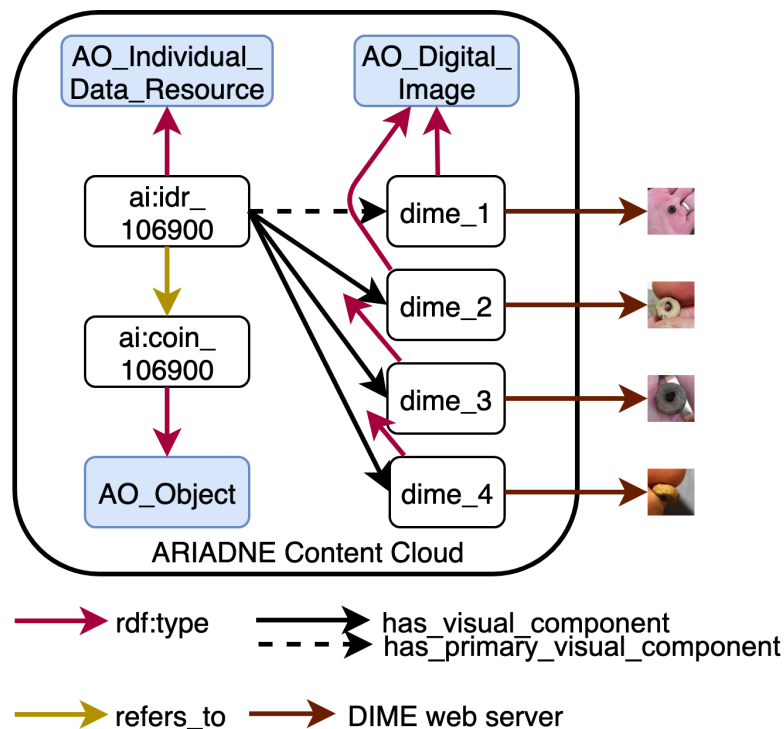


Figure 7.2: Model for images referred to in the ARIADNE Cloud.

7.2.3 Services

The class `AO_Service` has services as instances. Following the Parthenos Entity Model definition ([PEM Specifications 3.1](#)), a service is “an offer by some actor of their willingness and ability to execute an activity or series of activities upon request”. The descriptive properties defined on services are:

- `is_accessible_at`
- `has_functionality`
- `has_consumed_media`
- `has_produced_media`
- `has_consumed_format`
- `has_produced_format`
- `has_supported_language`
- `Has_technical_support`.

These extensions ensure that the AO-Cat ontology can accurately describe the heterogeneous data and services involved in ATRIUM’s demonstrators, and the ontology will continue to evolve throughout the project as new demonstrator requirements and usage patterns emerge, so that it remains aligned with both partners’ needs and best practice in semantic interoperability.

8. Updates to the portal

This section describes the updates and improvements implemented in the [ARIADNE Portal](#) since the beginning of the ATRIUM project. The work has focused on enhancing usability, improving search capabilities, supporting the deployment of the majority of the WP5 demonstrators within the portal, and ensuring better user experience for researchers accessing archaeological datasets.

Logo update

The ARIADNE Portal now features a new logo that clearly identifies the platform as the 'Ariadne Research Infrastructure'. This update more clearly aligns the portal with the overall branding of the ARIADNE RI.

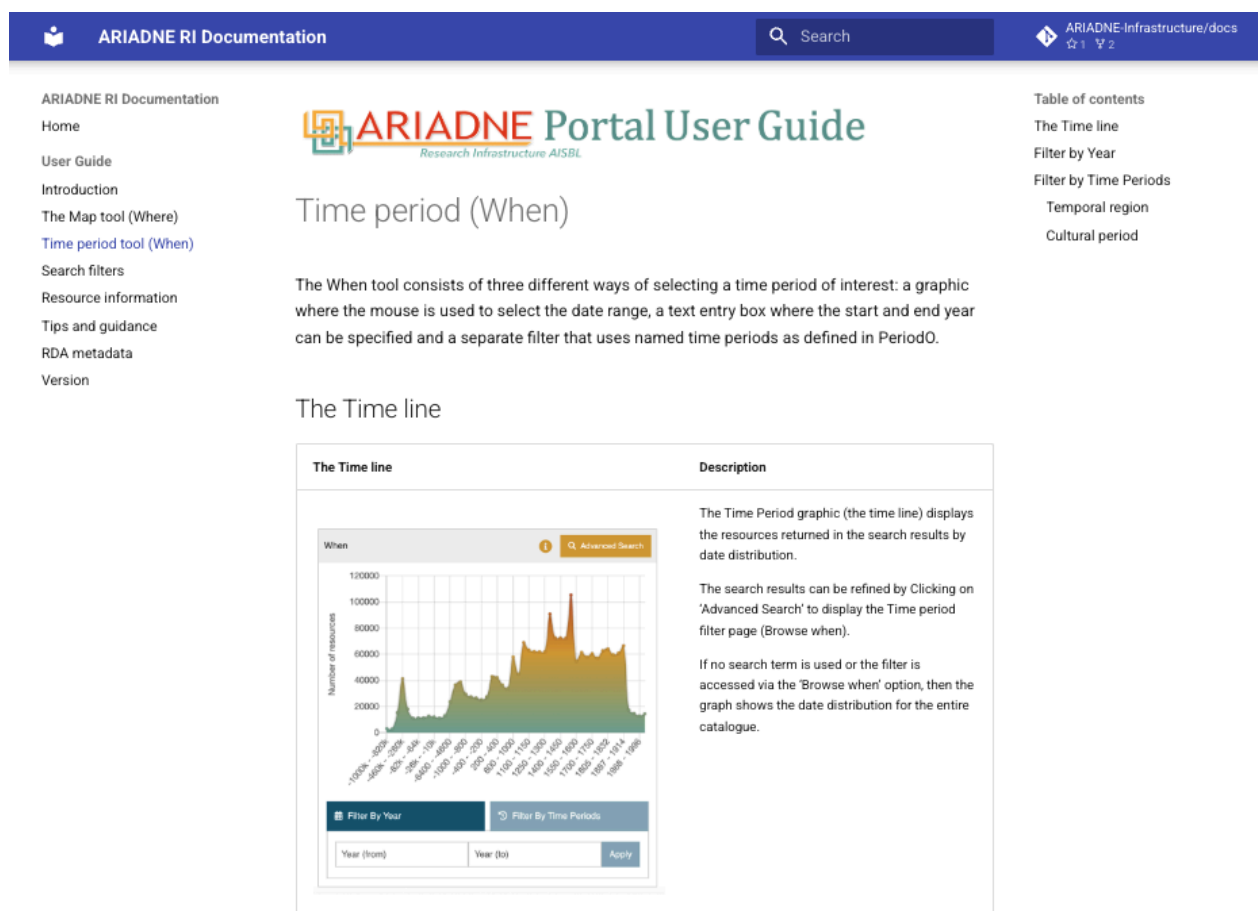


Figure 8.1: ARIADNE Portal homepage.

Improvements to the ARIADNE portal user manual

The original User Manual was written as a PDF Document. However, this was not considered user-friendly compared to an online version, which could be tailored to be more easily accessed and searched for information relevant to the particular topic of interest to the user. After assessing both a standard website editor (Wordpress) and GitHub documents, the latter was considered the best option in terms of sustainability and maintenance as well as being part of the ARIADNE-Infrastructure GitHub repository (as opposed to a stand-alone document).

The [new manual](#) was created by transferring the content and updating it as necessary for the latest updates to the Portal. The content has been divided into six sections so that each section can be accessed from the relevant context, e.g. the Resource section relates to the individual page options for a resource and the Time period (When) section (shown below) to the Map filter and Browse/When menu option. The footer's 'Portal Guide' link has also been updated to direct users to the comprehensive, extended user guide.



ARIADNE RI Documentation Search ARIADNE-Infrastructure/docs

ARIADNE RI Documentation
Home
User Guide
Introduction
The Map tool (Where)
Time period tool (When)
Search filters
Resource information
Tips and guidance
RDA metadata
Version

ARIADNE Portal User Guide
Research Infrastructure AISBL

Time period (When)

The When tool consists of three different ways of selecting a time period of interest: a graphic where the mouse is used to select the date range, a text entry box where the start and end year can be specified and a separate filter that uses named time periods as defined in PeriodO.

The Time line


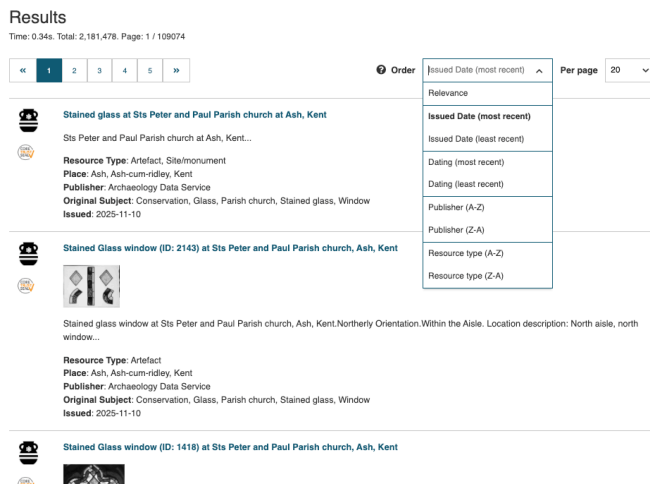
The Time line	Description
	<p>The Time Period graphic (the time line) displays the resources returned in the search results by date distribution.</p> <p>The search results can be refined by Clicking on 'Advanced Search' to display the Time period filter page (Browse when).</p> <p>If no search term is used or the filter is accessed via the 'Browse when' option, then the graph shows the date distribution for the entire catalogue.</p>

Figure 8.2: Screenshot of the online Portal User Manual

The menu on the left-hand side of the page shows the different sections and the menu on the right shows the sub-sections within the currently displayed page. This layout makes it quick and easy to find topics of interest.

This format also enables links to video such as the [online demonstration](#) about the portal given by Julian Richards. Work is in progress to create some shorter clips that can be embedded in the appropriate sections.

New sort options for the search



New sorting options have been added to the search results page, offering users greater control over how they view search results. Users can now sort results by relevance, issue date, dating, publisher, and resource type. This feature improves the discovery process, allowing researchers to sort search results according to their specific needs.

Figure 8.3: New sorting options in the search results.

Tile layers to all map searches

The map functionality within the portal now supports multiple tile layers, allowing users to switch between different base maps such as OpenStreetMap, Google Satellite, Terrain, Street, and Hybrid views. This feature enhances the geographic browsing experience, enabling researchers to contextualise resources within various geographical and topographical features.



Figure 8.4: The ARIADNE map viewer display

New Data Type search filter

A new data type search filter has been implemented, enabling users to limit their searches to specific data types such as Still image, Text, 3D and more. This improvement enhances the precision of searches for users interested in specific categories of data. It will enable findability of specific ATRIUM demonstrators, by allowing users to filter for all 3D, GIS or Sound datasets, for example.

▼ Data Type	
Enter text to filter on Data Types.	
Name	Hits ▼
Structured data	1304306
Still image	1112443
Text	207104
Geospatial	5193
Cad	1221
3d	822
Numeric	314
Video	63
Other	50
Software	4
Audio	4

Figure 8.5: The data type filter in the ARIADNE portal

Country filter

A country-based filtering option has been introduced, allowing users to filter on datasets covering particular countries.

New resource type 'E-publication'

A new resource type, 'E-publication', has been added to the catalogue, accompanied by a corresponding icon. This allows users to discover e-publications hosted at partner organisations. This category will encompass the scanned back runs of published journals for which metadata is being enhanced in T5.1.

Help text hover popups

Contextual help text popups have been added to the search field and sorting options. Users can now hover over an information icon to access quick guidance about how the search field filters or 'order by' feature works.

Added 'Search this area' to the minimap

A new interactive function, 'Search this area', has been added to the minimap within resource and search views. Users can now refine search results dynamically based on the area visible in the minimap interface.

Text, bug fixes, CSS & security updates

Multiple interface text updates, layout adjustments, and style (CSS) refinements were implemented across the portal. In parallel, backend maintenance included the patching of security dependencies and general stability and performance improvements.

New 'OR filter' logic on search page

The search interface is being expanded to support a more flexible logical filtering system. The upcoming 'OR filter' functionality will allow users to apply AND logic between different filter categories and OR logic within a single category. This improvement will give users a new way to dynamically enhance their search results.

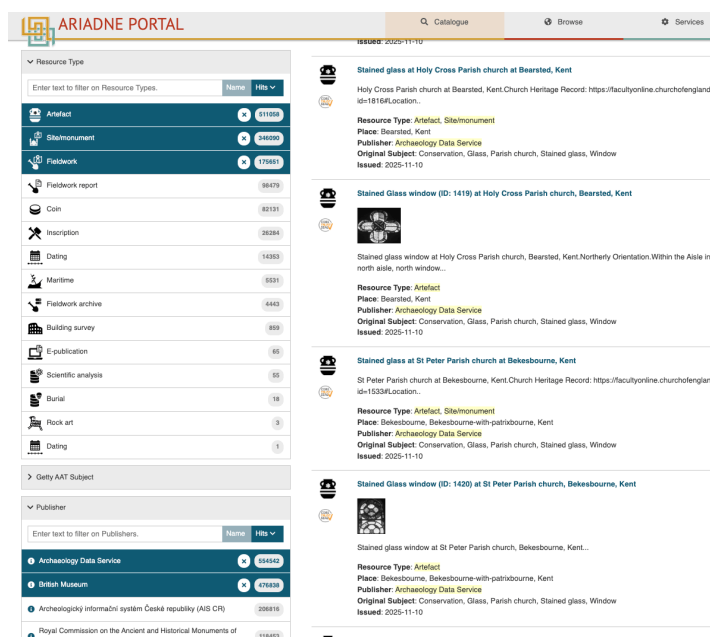


Figure 8.6: Filter showing multiple options selected as part of an OR database join.

Enhanced search logic and improved autocomplete

The search engine now makes use of OpenSearch's 'simple_query_string' functionality, allowing more sophisticated parsing of user queries and significantly improving search relevance, leading to more accurate search results. The autocomplete feature has also been upgraded, ensuring that users receive more meaningful suggestions as they type.

Displaying digital media

The portal was updated to support a new 'Digital Media' field, enabling the integration and display of audio, video, and 3D digital resources in the landing pages of records. This enhancement broadens the portal's scope, enabling interactive preview and exploration of multimedia content.

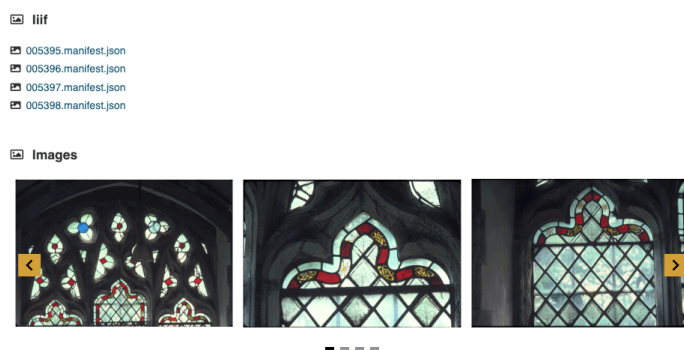


Figure 8.7: IIIF Manifest and Image display in the Portal.


9. Conclusion

ATRIUM delivers reusable research workflows and associated demonstrators for five key data types in archaeology (text, images, 3D, sound, geospatial). This report focuses on the demonstrators that illustrate the workflows with real use-cases for archaeology.

Despite a few setbacks in some subtasks, tasks have generally progressed at pace and sometimes also ahead of schedule. Some major successes can be highlighted:

- NLP pipelines for the extraction of subject, place, and time were improved and refined drastically.
- A set of advanced media viewers have been deployed with various partners.
- Several partners have deployed OAI-PMH APIs to facilitate the automated aggregation of datasets metadata.
- The AO-Cat ontology and the ARIADNE portal are systematically updated, closely following the requirements of the project.
- Case studies are well-defined and well-rounded, some of them are close to completion.

In the next months, the Work Package will scale up enrichment work (text and images), complete the case studies and finish disseminating the results in the ARIADNE portal and through additional dissemination outputs.

Besides this report, the Work Package has also shared our progress internally in a two-hour presentation-based general meeting of WP5 on November 26. The slides of this event are available here:  WP5_general_meeting_20251126_slides.pdf . We expect to organise a similar meeting at month 42 to wrap up the activities of the WP.

We are looking forward to deploying the final version of all of our demonstrators and ensuring selected providers' data and case studies are available, enriched, findable, and cross-searchable in the ARIADNE portal.

References

Andersson Palm, Lennart. *The database Sweden 1570–1810: population, agriculture, land ownership – Agricultural statistics on parish level, 1810 (1.0)*. Data set. University of Gothenburg, 2014. <https://doi.org/10.5878/002159>.

Binding, Ceri. n.d. "Rematch2: Github Open Source Pipeline." GitHub. Accessed November 2025. <https://github.com/cbinding/rematch2>.

Binding, Ceri, and Douglas Tudhope. 2023. "Automatic Normalization of Temporal Expressions." *Journal of Computer Applications in Archaeology* 6, no. 1: 24–39. <https://doi.org/10.5334/jcaa.105>.

Binding, Ceri, and Douglas Tudhope. 2024. "KOS-based enrichment of archaeological fieldwork reports." *Knowledge Organization* 51, no. 5 (2024): 292–299. <https://doi.org/10.5771/0943-7444-2024-5-292>.

Binding, Ceri, and Douglas Tudhope. 2025a. "Go with the flow – workflows as a recipe for reproducible results." Presentation at CAA 2025, Athens, May 2025. <https://pure.southwales.ac.uk/en/activities/go-with-the-flow-workflows-as-a-recipe-for-reproducible-results/>.

Binding, Ceri, and Douglas Tudhope. 2025b. "Open-source tools for archaeological temporal expressions." Presentation at CAA 2025, Athens, May 2025. <https://pure.southwales.ac.uk/en/activities/open-source-tools-for-archaeological-temporal-expressions/>.

Clérice, Thibault, Juliette Janès, Hugo Scheithauer, et al. 2024. "Diachronic Document Dataset for Semantic Layout Analysis." Preprint, HAL, November 14. <https://hal.science/hal-04784161>.

Clérice, Thibault, Juliette Janès, and Sarah Bénérière. 2025. "Automatic Text Recognition using Object Detection with eScriptorium." Workflow, SSHOMP, last modified November 28. <https://marketplace.sshopencloud.eu/workflow/sS4gSB>.

Faka, M., Orabi, R., Tsagka, A., Papageorgiou, A., Vassallo, V., Hermon, S., Bakirtzis, N. (2025), Hypothetical Reconstruction for the Conservation, Preservation and Valorisation of Cultural Heritage: the Kampanopetra Basilica in Salamis, Cyprus. In *Digital Heritage*, Campana, S., Ferdani, D. Graf, H. Guidi, G. Hegarty, Z., Pescarin, S., Remondino, F. (eds), {The Eurographics Association, 10.2312/dh.20253309

High Speed Two Ltd., and MOLA Headland Infrastructure. *Digital Archive from Archaeological Excavations at St James's Gardens Burial Grounds, Euston, Camden, Greater London, 2018-2022 (HS2 Phase One)*. Data set. York: Archaeology Data Service, 2025. <https://doi.org/10.5284/1122289>.

High Speed Two Ltd., and MOLA Northampton. *Data from Stories of St James's Burial Ground, Euston, Camden, Greater London, 2021-2023 (HS2 Phase One)*. Data set. York: Archaeology Data Service, 2023. <https://doi.org/10.5284/1118189>.

Historic England. "Archaeological and Cultural Periods." PeriodO Authority. Accessed November 2025.
<https://client.perio.do/?page=authority-view&backendID=web-https%3A%2F%2Fdata.perio.do%2F&authorityID=p0kh9ds>.

Janès, Juliette, Sarah Bènière, Lucence Ing, and Thibault Clérice. 2025. "LADaS Annotation Guidelines." Preprint, HAL, September 12.
<https://inria.hal.science/hal-05252327>.

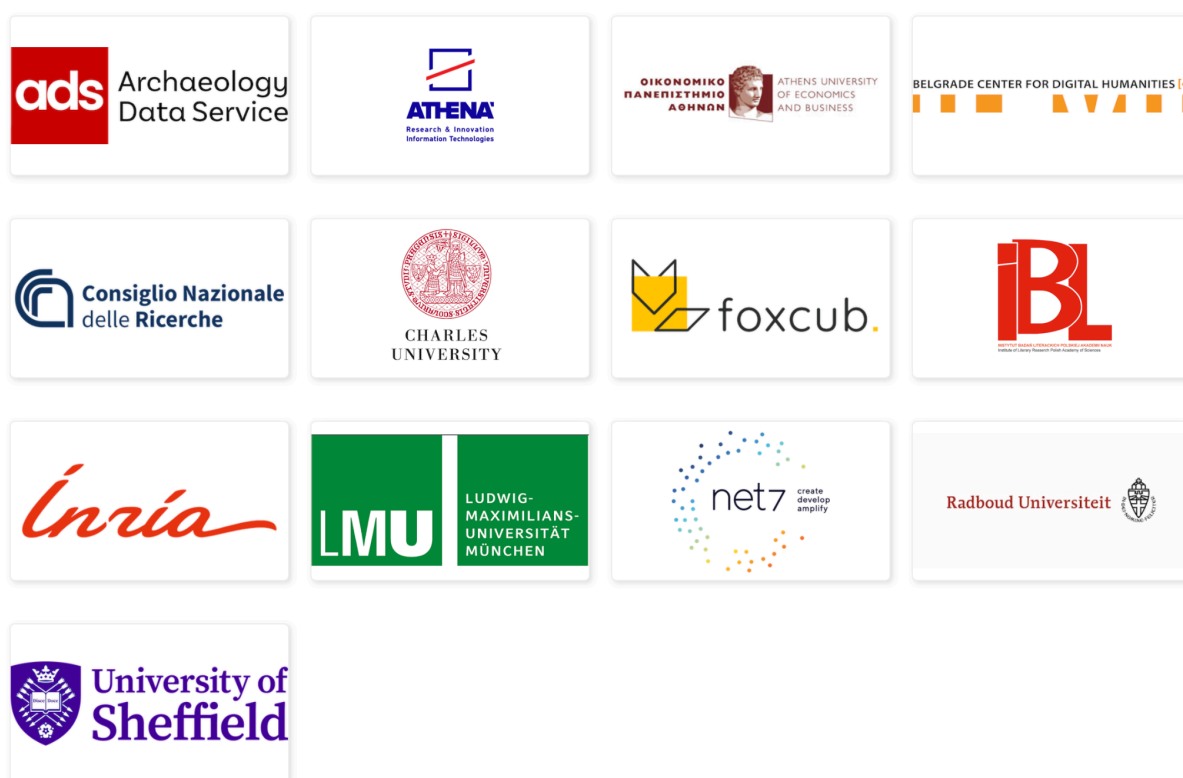
Roux G.: La basilique de la Campanopétra, vol. 15. Salamine de Chypre, 1998. URL:
https://www.persee.fr/doc/salam_0000-0000_1998_arc_15_1. 2, 4, 5

Consortium

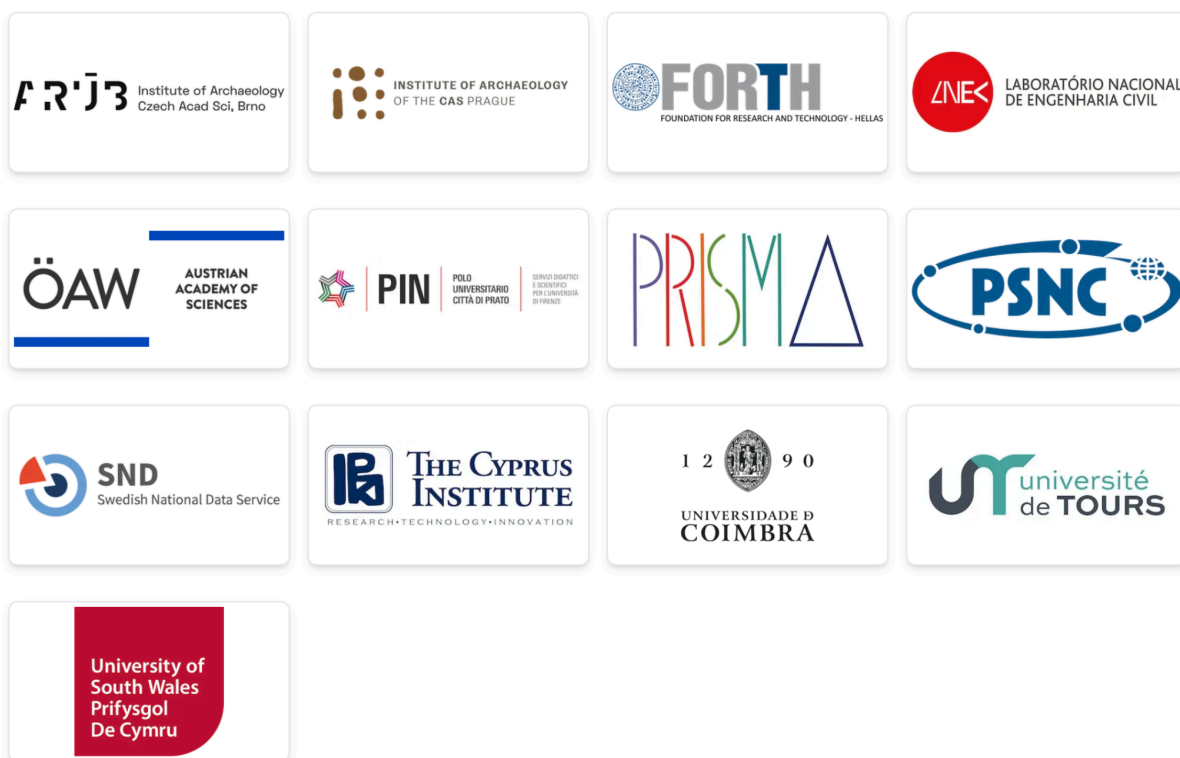
Research Infrastructures



Beneficiaries



Affiliated Entities



Disclaimer

All information provided reflects the status of the ATRIUM project at the time of writing and may be subject to change. Neither the ATRIUM Consortium as a whole, nor any single party within the ATRIUM Consortium warrant that the information contained in this document is capable of use, nor that the use of such information is free from risk. Neither the ATRIUM Consortium as a whole, nor any single party within the ATRIUM Consortium accepts any liability for loss or damage suffered by any person using the information.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. Funded by the European Union. Grant Agreement number 101132163. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

© 2025 by the authors, the ATRIUM consortium. This work is licensed under a "CC BY 4.0" license.



**Funded by
the European Union**