Trustworthy AI-based Multi-Modal Data Processing for Cultural Heritage Monitoring and Preservation

Lingxiao Kong[1], Athos Agapiou[2], George Pavlidis[3], Zeyd Boukhers[1]

[1]Fraunhofer Institute for Applied Information Technology FIT, [2]Cyprus University of Technology, [3]ATHENA Research Center

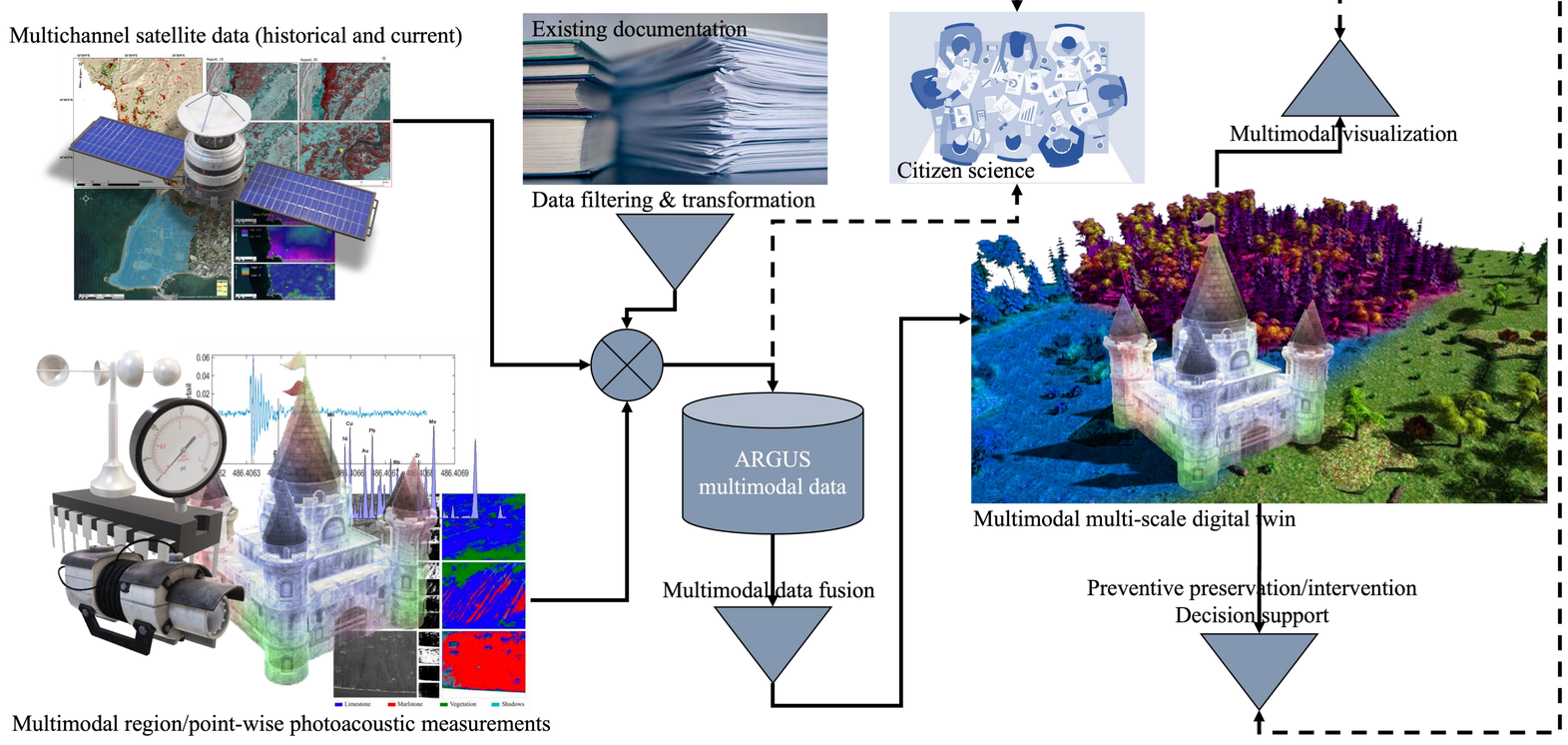
1. CH Data Challenges

Heterogeneous data of diverse types and modalities are continuously fed into the ARGUS digital twin, requiring sophisticated processing before they can be effectively utilized. Key challenges include:

- Inconsistent formats and non-interoperable standards
- Missing or ambiguous metadata
- Limited spatial and temporal coverage
- Incompatibility for multi-modal AI analysis
- Misalignment with FAIR principles

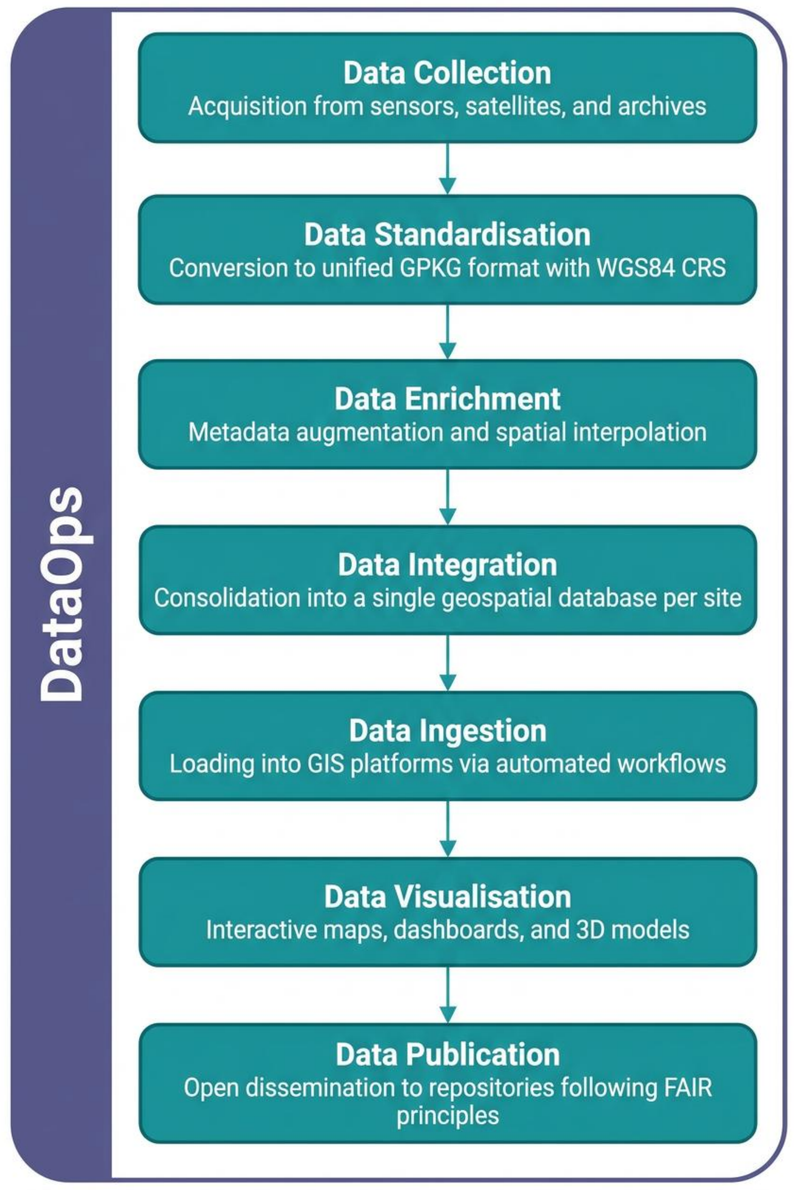
2. Data Processing Pipeline

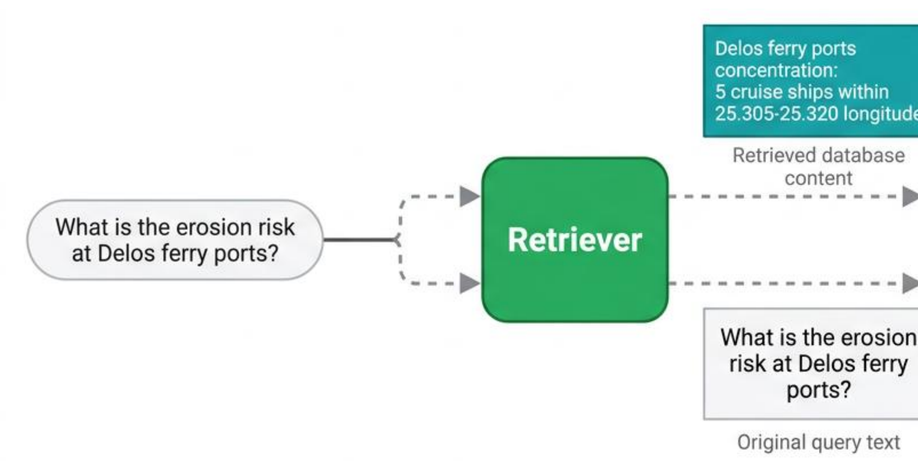
The CH data processing pipeline comprising **seven sequential stages** from data sources through AI-ready data, governed by an overarching DataOps framework. Each stage progressively transforms raw heterogeneous data spanning geometric, visual, and documentary modalities into standardized, enriched, and queryable resources, with DataOps ensuring automation, monitoring, and continuous quality control throughout the entire lifecycle, providing a reliable data foundation for AI-driven heritage monitoring and preservation.

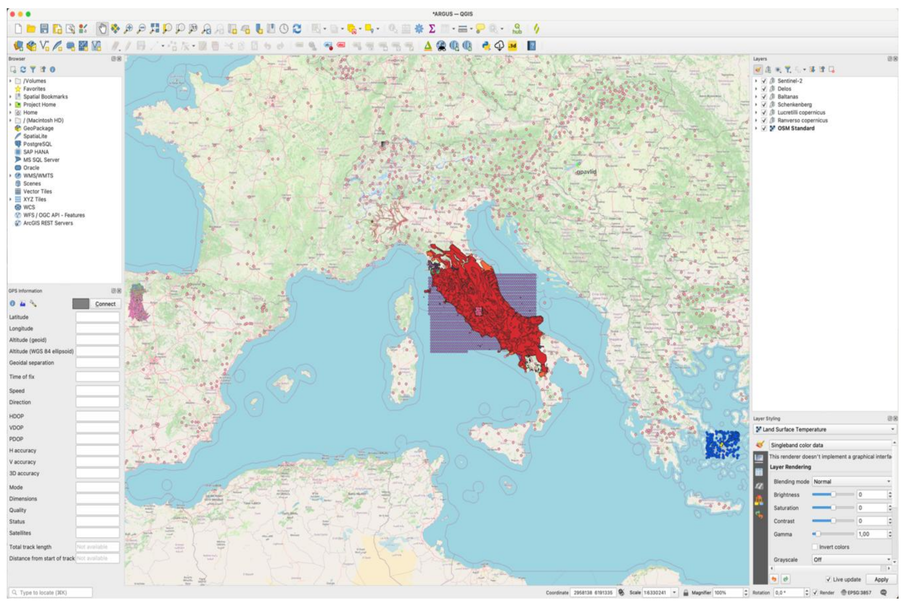


3. RAG-based LLM Application

The application supports two functionalities. For **database querying**, natural language queries are fed into a Retriever that retrieves relevant records from the internal GPKG database, which are jointly provided to the Language Model to generate grounded, contextually accurate responses. For **attribute annotation**, the same architecture retrieves definitions and unit references from external knowledge bases such as WikiData to generate missing attribute descriptions. All outputs are labelled [GENERATED] with source citations to ensure transparency and traceability.







DATASETS INTEGRATED
249 → 1
ADF, CSV, GDB, GPKG, SHP, TIF, XLS, XLSX

PIPELINE SUCCESS RATE
98.8%
246 of 249 succeeded

PILOT SITES COVERED
5
Schenkenberg, Lucretili, Delos, Ranverso, Baltanás

OUTPUT FORMAT
GPKG/WGS84 CRS

TEMPORAL COVERAGE
days → sec
Nearest-neighbour imputation

SPATIAL COVERAGE
22% → 76%
Nearest-neighbour imputation

ENRICHED METADATA
14% → 100%
Descriptions and units added

QUERY EFFICIENCY
hours → minutes
LLM-supported querying

4. Processing Results and Further Steps

QGIS visualization of the integrated ARGUS geospatial database and key processing results, demonstrating that the proposed framework effectively transforms heterogeneous cultural heritage data into a standardized, enriched, and queryable resource, providing a reliable foundation for **downstream AI solutions** including anomaly detection, visualization, and decision support for heritage monitoring and preservation.

This work has been supported by the ARGUS EU project (Grant Agreement No. 101132308), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or of the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Dr. Zeyd Boukhers
Email: zeyd.Boukhers@fit.fraunhofer.de
Web: https://fit.fraunhofer.de/fdda