Dür.air: reconciling acquisition and interpretation in cultural heritage field documentation

Livio De Luca, Florent Comte, Anthony Pamart

MAP – UPR 2002 CNRS, Campus CNRS, 31 chemin Joseph Aiguier, 13009 Marseille, France
(livio.deluca, florent.comte, anthony.pamart)@map.cnrs.fr

Keywords: Persistent augmented reality; field documentation; reality-based annotation; interpretation-driven workflows.

Abstract

Field-based scientific and professional activities involved in the study and conservation of cultural heritage rely on observation practices that are inherently spatial, temporal, and interpretative. However, contemporary 3D digitization workflows increasingly dissociate data acquisition from interpretation, favouring exhaustive capture in the field followed by deferred analysis. This separation often leads to a loss of observational intent, reduced semantic coherence, and limited support for multi-temporal field studies. This paper presents *dür.air*, a mobile augmented reality application designed to reconnect acquisition, annotation, and interpretation directly on site. The proposed approach is structured around three methodological pillars. First, *persistent spatial anchoring* enables annotations to remain coherently aligned across multiple field sessions through augmented reality relocalisation. Second, a *spatio-temporal semantic enrichment* framework supports both synchronous (in situ) annotation and asynchronous (ex situ) interpretation, with consistent projection and reprojection between 2D images, depth data, and photogrammetric 3D models. Third, an *interpretation-driven and frugal acquisition* strategy explicitly links each capture to a scientific observation, reducing over-acquisition while increasing semantic density.
Implemented as a fully mobile and autonomous system operating on a single handheld device, *dür.air* integrates augmented reality tracking, semantic annotation, and on-device photogrammetric reconstruction into a unified workflow. Rather than treating interpretation as a post-processing step, the system embeds it within the acquisition process itself, enabling cumulative, multi-temporal documentation. This work defines a methodological framework for field documentation in which augmented reality is used not only for visualisation, but as a medium for preserving scientific meaning in space and time.

1. Introduction

1.1. Motivation and problem statement

Field-based disciplines involved in the study and conservation of cultural heritage rely on in situ observation practices that are inherently spatial, temporal, and interpretative (Hodder, 1999; Lucas, 2012; Ingold, 2011). In archaeology, architectural history, conservation–restoration, and landscape studies, observation, recording, and interpretation have traditionally been tightly coupled within a single *situated act*, materialised through annotated drawings, measured sketches, stratigraphic logs, and field notebooks produced directly on site (Courty et al., 1989; Letellier, Schmid and LeBlanc, 2007; Brandi, 1963; Corner, 1999).
With the widespread adoption of 3D digitisation, documentation workflows have shifted toward data-intensive acquisition followed by deferred processing and interpretation. While these approaches provide geometrically accurate representations, they introduce a semantic gap between acquisition and interpretation: observational intent and contextual knowledge formulated in situ are often weakened (or lost) when analysis is displaced in time and space.
This limitation is particularly critical in multi-temporal field studies, where observations made during successive visits must remain spatially coherent and semantically comparable. Yet current 3D workflows provide limited support for persistent, spatially anchored annotations that can be reactivated and enriched across sessions.

The central question addressed in this paper is therefore:

how can data acquisition, annotation, and interpretation be reconnected directly on site, while preserving spatial and temporal coherence across multiple field sessions within a mobile and autonomous system?

2. Scientific positioning

Our work approaches spatial documentation as an interpretative process rather than as a neutral recording of reality. Acquisition is thus considered a selective and intentional act, guided by hypotheses and disciplinary perspectives, in which each captured datum is explicitly linked to the observation and interpretation that motivated it.
Augmented reality (AR) provides a particularly suitable framework for this approach, as it is grounded in simultaneous localisation and mapping (SLAM), which estimates camera motion while incrementally reconstructing a spatial model of the environment (Klein and Murray, 2007). This enables spatially anchored annotation, real-time feedback, and persistent alignment with the physical scene. Observations can therefore be recorded at the very place where they are formulated and subsequently revisited across field sessions, while preserving spatial and temporal continuity.
In this perspective, 3D digitisation is not conceived as an end product, but as a geometric support for the consolidation and asynchronous enrichment of in situ observations. From a technological standpoint, the work focuses on the integration of persistent AR tracking, spatio-temporal annotation, and mobile photogrammetry into a unified and autonomous field documentation workflow operating on a single mobile device.

2. Related work

2.1. Field-based 3D documentation

Photogrammetry, terrestrial laser scanning, and, more recently, mobile depth-sensing and LiDAR technologies have become standard tools for documenting cultural heritage objects and sites, enabling the production of geometrically accurate 3D representations at multiple scales (Remondino and Campana, 2014; Sapirstein, 2016; Teppati Losè et al., 2022; Kędziorski et al., 2024; Antón et al., 2025). Contemporary workflows

typically combine extensive field acquisition with off-site processing stages such as registration, structure-from-motion, dense reconstruction, meshing, and texturing. These stages, often involving heavy computation, multiple software environments, and transformations between geometric representations (e.g. point clouds, meshes, images, vector derivatives), produce the models that subsequently serve as supports for observation, annotation, and analysis, and may be carried out by different actors with distinct expertise.

While such pipelines provide high geometric fidelity, they tend to dissociate acquisition from interpretation, which is deferred and progressively detached from the original spatial, temporal, and cognitive context of observation. It is now widely acknowledged in scientific and professional communities that this separation limits the effectiveness of field-based documentation, particularly in multi-temporal contexts involving repeated site visits, longitudinal observation, or fine-grained semantic analysis. Large data volumes further entail high processing costs and a weak formalisation of observational intent, as the hypotheses and interpretative choices guiding acquisition are rarely captured at the moment of recording.

In multi-temporal documentation, current approaches generally rely on the comparison of independently produced datasets acquired at different times, requiring complex alignment procedures and substantial manual interpretation. Although temporal metadata are usually available, few systems explicitly manage observation sessions as a structuring principle for managing semantic entities over time (Dell'Unto et al., 2017). Temporal continuity is therefore most often reconstructed retrospectively rather than embedded in the documentation process itself, reinforcing workflows in which geometric capture is driven by technological constraints while scientific interpretation and semantic context remain weakly integrated.

2.2. Semantic enrichment of spatial data

Semantic enrichment of spatial data has been widely investigated in the field of cultural heritage, notably for the annotation of images, point clouds, and 3D models (Ponchio et al., 2020), and is most often carried out as a post-processing activity on reconstructed models or orthophotos in desktop or web-based environments (Abergel et al., 2023). In parallel, research in Augmented Reality has explored persistent, spatially anchored AR experiences that remain accessible and modifiable over time and across users, highlighting the importance of long-term spatial coherence in collaborative workflows (Guo et al., 2019). Despite their effectiveness for detailed analysis and visualisation, these approaches introduce a temporal and cognitive gap between observation and annotation: the delays induced by acquisition, reconstruction, and visualisation pipelines may lead to a partial loss of the contextual and perceptual information available at the moment of in situ observation. Augmented reality has therefore attracted increasing attention for cultural heritage documentation, measurement, and visualisation, as it enables spatially situated interaction, real-time tracking, and on-site visual feedback (Bekele et al., 2018). However, most AR-based applications remain primarily oriented toward visualisation or public dissemination rather than toward scientific documentation. Persistent spatial anchoring is often limited to single sessions, and robust mechanisms for relocalisation, multi-session reuse, and integration with post-processing and 3D reconstruction environments remain limited. More generally, whether in desktop, web-based, or AR-based systems, semantic information is still frequently treated as an additional layer attached to geometric data rather than as an integral component of the acquisition process itself. While such semantic annotation frameworks allow the accumulation of multiple interpretative layers on a same 3D representation, they are most often produced a posteriori and remain weakly connected to the temporal dynamics of field observation. In practice, interpretative acts occur at a much finer and more frequent rhythm than heavy 3D acquisition campaigns, so that spatio-temporal continuity between observation, geometry, and semantics is rarely maintained. As a consequence, annotations often lack an explicit link to the situated observational and interpretative acts that motivated their creation.

3. Proposed approach

3.1. Reconciling data acquisition and interpretation

The proposed approach aims to overcome the structural separation between acquisition, interpretation, and temporal reuse that characterises most current 3D documentation workflows. Rather than treating geometry capture, semantic enrichment, and visualisation as sequential or loosely coupled stages, these processes are integrated into a single field-based workflow explicitly designed to support multi-temporal investigation.

Building on previous contributions in spatialised 2D/3D annotation (Manuel et al., 2014), SLAM-based augmented reality for in situ 3D annotation (Abergel et al., 2019), and image-based multimodal registration and photogrammetric reconstruction (Pamart et al., 2020), *dür.air* combines persistent augmented-reality relocalisation, spatio-temporally anchored annotation, and mobile photogrammetric reconstruction. This integration addresses several key limitations identified in prior work: the loss of interpretative context between in situ observation and ex situ analysis, the difficulty of maintaining spatial coherence across repeated site visits, and the fragmentation between situated perception, semantic practices, and 3D reconstruction pipelines.

Semantic information is captured at the moment it is formulated, spatially and temporally anchored, and preserved across sessions through persistent alignment. Augmented reality is therefore considered not merely as a visualisation interface, but as an operational medium for embedding scientific meaning directly into spatial data and for ensuring continuity between observation, interpretation, and subsequent analysis.

3.2. Three complementary pillars

The approach is structured around three complementary methodological pillars:

- *Persistent spatial anchoring* (section 4), ensuring that observations and annotations remain spatially coherent and reusable across sessions;
- *Spatio-temporal semantic enrichment* (section 5), allowing interpretations to evolve while preserving geometric and contextual consistency;
- *Interpretation-driven acquisition* (section 6), in which capture is guided by scientific intent, reducing redundancy while increasing semantic density.

Together, these principles define field documentation as a continuous and cumulative process in which observation, interpretation, and memorisation are tightly coupled, within an open and evolutive framework in which spatial registrations, projections, and models remain recalculable and interoperable with established disciplinary workflows.

3.3. System-level implementation

These methodological principles are implemented in a unified mobile architecture. *Dür.air* is currently developed as a native application for Apple iPad and organises projects as persistent spatio-temporal entities integrating AR spatial maps, image and depth data, photogrammetric models, camera paths, and

structured annotations. The internal data model supports multi-session alignment, versioning, and offline-first operation, while remaining compatible with external platforms and data sharing infrastructures such as *aïoli* (Abergel et al., 2013) and *quasi.modo* (manuscript under review). This organisation enables the same spatio-temporal and semantic information to be consistently exploited both in situ, within augmented reality, and ex situ, within on-device analytical environments, while being designed for subsequent integration into collaborative and web-based platforms, without fragmenting the underlying representations.

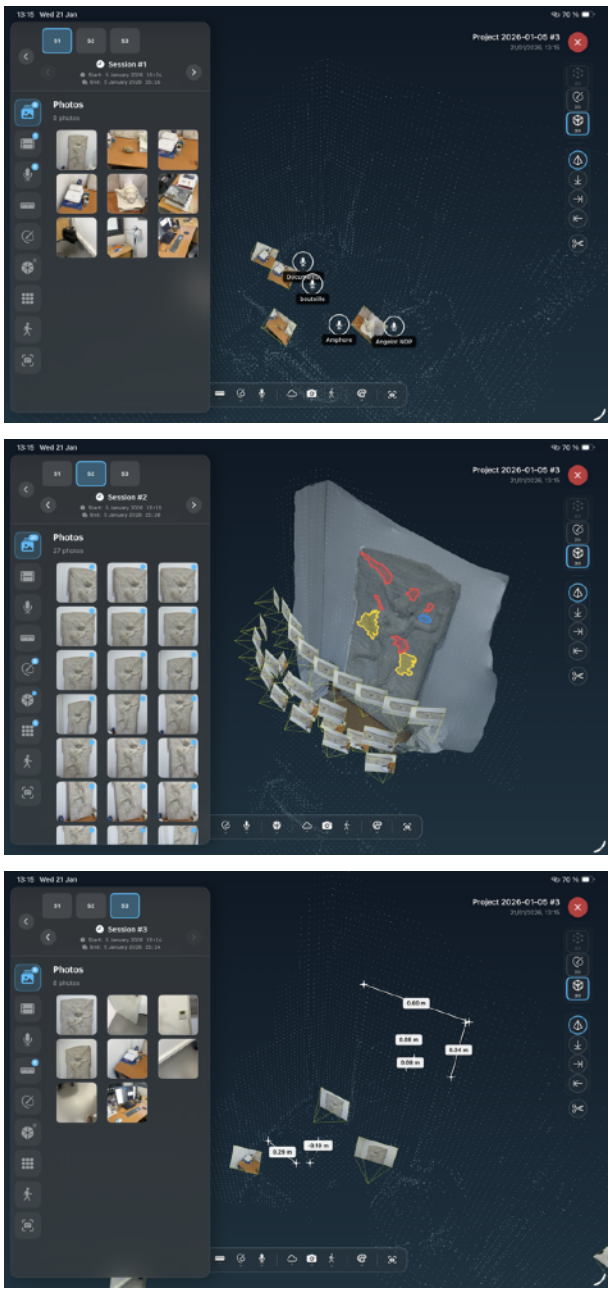

Figure 1. Multi-temporal field sessions aligned in a persistent AR reference frame: voice annotations (top), photogrammetric reconstruction and semantic annotation (middle), and in situ measurements (bottom)

4. Pillar 1: Persistent Spatial Anchoring

4.1. Spatial persistence for diachronic analysis

Scientific field documentation is intrinsically multi-temporal: observations are produced during successive site visits, often separated by long intervals, and must be compared, reinterpreted, and enriched within a stable spatial reference in order to support cumulative analysis, diachronic reasoning, and monitoring. Yet most current mobile acquisition and annotation systems provide only session-bound spatial coherence, which prevents observations from being reliably reactivated and contextualised over time.

The first pillar of the proposed approach therefore addresses the problem of persistent spatial anchoring: how to ensure that observations, annotations, and measurements remain spatially coherent and reusable across multiple field sessions, despite variations in viewpoints, partial scene changes, and tracking drift. Our approach conceives fieldwork as a sequence of temporally indexed observation sessions sharing a progressively refined and persistent spatial reference.

This pillar relies on a combination of persistent AR relocalisation, versioned spatial memory, distributed anchoring, and drift correction, enabling environmental maps, spatial entities, and observation sessions to be reused, incrementally refined, and coherently realigned across time. By explicitly modelling multi-temporal sessions within a common reference frame, annotations, measurements, and derived products acquired during successive field visits can be spatially superimposed, compared, and reinterpreted.

Together, these mechanisms transform spatial persistence from a short-lived tracking capability into a long-term, cumulative spatial memory supporting diachronic comparison, progressive enrichment of observations, and the continuity of spatial reasoning.

4.2. Persistent AR relocalisation and World Map versioning

Spatial persistence is achieved through the use of ARKit's ARWorldMap, which encodes a sparse three-dimensional representation of the environment based on visual features and, when available, LiDAR depth information. Instead of relying on a single snapshot of the environment, *dür.air* introduces a versioned World Map model in which each field session produces its own spatial map, stored together with quality metrics and lineage information.

For each session, a WorldMapVersion object records:

- the number and spatial distribution of ARKit feature points,
- the spatial extent and centroid of the reconstructed environment,
- links to previous versions from which the map was incrementally improved,
- and an automatically derived quality class (e.g. excellent, good, fair, poor).

This design results in a temporal graph of spatial references rather than a flat collection of maps. During relocalisation, the system can select the most appropriate map according to both geometric quality and contextual proximity, allowing robust alignment even when lighting conditions, viewpoints, or partial scene changes differ between visits.

From a methodological point of view, this versioned spatial memory supports the notion of a progressively refined, cumulative spatial reference rather than a static baseline.

4.3. Frame anchor constellation and drift correction

While ARKit provides global relocalisation, local drift may still affect the spatial consistency of previously recorded observations when a session is resumed. To compensate for this effect, *dür.air* implements a dedicated frame-anchoring strategy based on the distribution of multiple spatial anchors for each key frame, for instance at the time an annotation is created.
For each such frame, a constellation of anchors is generated and distributed over a regular image grid and across multiple depth layers. Anchor placement is weighted by the local density of ARKit feature points in order to favour geometrically stable regions of the scene. Rather than relying on a single centroid-based anchor, this multi-point configuration captures the spatial structure of the environment at different scales.
Upon relocalisation, the current positions of these anchors are matched to their original coordinates. A robust pose correction is then estimated by computing the rigid transformation (rotation and translation) that minimises the residual error over all inlier anchors. This transformation is applied to the camera pose and to all associated spatial entities, yielding a coherent realignment of the observation space. In practice, this mechanism preserves the relative geometry between annotations, measurements, and reconstructed surfaces, and locally corrects accumulated tracking drift in areas already covered by key frames from previous sessions.

4.4. Multi-temporal sessions

Beyond static spatial persistence, the system explicitly models fieldwork as a sequence of temporally indexed observation sessions, all registered within a common and stable spatial reference frame. Each session aggregates its own set of images, depth data, annotations, measurements, and derived products, while remaining geometrically aligned with those acquired during previous visits, within the accuracy limits imposed by the current device and programming environment.
This organisation enables the selective visualisation, comparison, and enrichment of observations according to their temporal context. Annotations and measurements created at different moments can be superimposed, filtered, and reinterpreted within the same three-dimensional reference, supporting cumulative documentation and diachronic analysis of the same physical areas or features (Fig. 1).
Beyond their immediate in situ alignment, the preserved images and associated depth maps also constitute a coherent spatio-temporal corpus that can later be re-registered with higher precision using external photogrammetric pipelines, allowing the spatial reference to be further refined in post-processing or in other computational environments. In this way, multi-temporality is treated not as a simple metadata attribute, but as an explicit organisational dimension of the spatial reference itself.

5. Pillar 2: Spatio-temporal semantic enrichment

5.1. From synchronous field annotation to asynchronous analytical enrichment

Scientific interpretation spans multiple temporalities and representational spaces, from in situ observation to later analysis on images and reconstructed 3D models. A central methodological challenge is therefore to support both *synchronous* and *asynchronous* semantic enrichment while preserving strict spatial and temporal coherence.
The second pillar addresses this challenge by modelling each observation as a spatio-temporal unit that tightly couples image space, three-dimensional geometry, and acquisition context. Each captured frame is associated with its camera pose, intrinsics, depth information, and timestamp, providing a common reference for projecting, reprojecting, and aligning annotations across 2D images, 3D space, photogrammetric models, and orthographic views. Annotations created in the field inherit this anchoring and can later be refined or extended without loss of geometric or temporal consistency.
Bidirectional projection between 2D and 3D representations allows semantic entities to circulate consistently across modalities—between images, spatial models, and metric orthographic views—so that interpretation can evolve over time while remaining grounded in a shared spatio-temporal reference.

5.2. 2D-3D bidirectional projection

The technical realisation of this pillar relies on a unified multi-modal projection and unprojection framework designed to maintain geometric and temporal consistency across heterogeneous sources of spatial information.
The system integrates complementary sensing and reconstruction modalities (RGB imagery, LiDAR depth, ARKit spatial mapping, and, when computed, dense photogrammetric meshes) within a single spatio-temporal reference model.
A hierarchical fallback strategy ensures robustness: when a high-resolution photogrammetric surface is available it is used as the primary geometric support; otherwise, LiDAR depth is exploited; and, in the absence of both, planar surfaces and feature-based geometry estimated by ARKit provide a coarse but spatially consistent reference. Each observation is represented as a spatio-temporal capture unit corresponding to a single frame of the acquisition stream (Fig. 2).
A frame associates:

- high-resolution RGB image,
- a registered LiDAR depth map,
- a per-pixel confidence map (depth reliability),
- the full calibrated camera model,
- a precise timestamp.

This multimodal bundle constitutes the minimal coherent support for semantic anchoring, projection, and reprojection across representations.

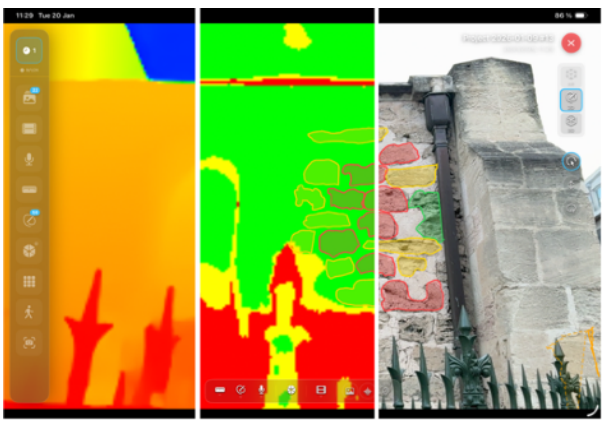

Figure 2. Multi-modal frame representation used for spatio-temporal anchoring and projection. From left to right: LiDAR depth map, confidence map, and corresponding RGB image with spatially anchored annotations.

5.3. Multi-source 2D–3D projection and visibility modelling on depth-based geometric layers

Annotations are created and visualised in real time by the user directly on the live camera frames during the AR session. For each frame, the system relies on the calibrated ARKit camera

model (intrinsics and pose), the associated LiDAR depth map, and its confidence map.
World-space coordinates of spatial entities are transformed into camera space and projected onto the image plane using the intrinsic projection matrix. These forward and backward 2D–3D–2D projections are implemented on iOS using a GPU-accelerated pipeline based on Metal, allowing all geometric transformations, depth tests, and confidence-weighted visibility computations to be performed in real time during acquisition (Fig. 3).
Visibility is evaluated on-the-fly by comparing the projected depth of each point with the corresponding LiDAR depth value, weighted by the confidence map. This depth-consistency test enables real-time classification of projected points as visible or occluded, preventing erroneous reprojection of spatial annotations onto foreground surfaces and allowing partial visibility to be handled explicitly.
Conversely, when the user delineates a region directly on an image, unprojection into three-dimensional space is first performed using LiDAR depth, yielding a set of metric 3D anchor points with associated confidence values. In areas where LiDAR data are missing or unreliable, the system automatically falls back to ARKit's estimated planar geometry and sparse feature-based depth, ensuring continuity of spatial anchoring under degraded sensing conditions.
This multi-level strategy guarantees that semantic regions drawn in real time can always be lifted from 2D to 3D using the best available geometric support, and subsequently reprojected consistently onto orthographic views or reconstructed models within a unified spatio-temporal reference frame.

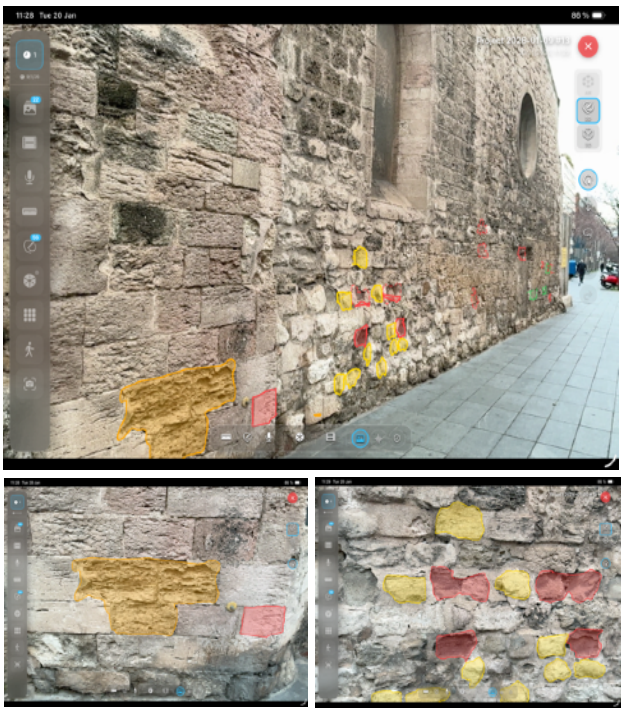

Figure 3. Real-time depth and confidence-aware reprojection of spatial annotations in the augmented reality view.

5.4. Photogrammetric geometry as a high-resolution and long-range anchoring layer

While LiDAR depth maps provide dense and reliable geometry in the near field, their effective range remains limited on current mobile devices (typically reliable to ~5 m on earlier iPad Pro models and up to ~10 m on the latest hardware), particularly for elevated architectural elements such as cornices, vaults, or upper façade. In addition, their spatial resolution is insufficient for fine-scale morphological analysis.
To overcome both the limited sensing range and the need for high geometric fidelity, *dür.air* integrates on-device photogrammetric reconstruction, computed through RealityKit and GPU-accelerated multi-view stereo, as a complementary and higher-resolution spatial support (Fig. 4).

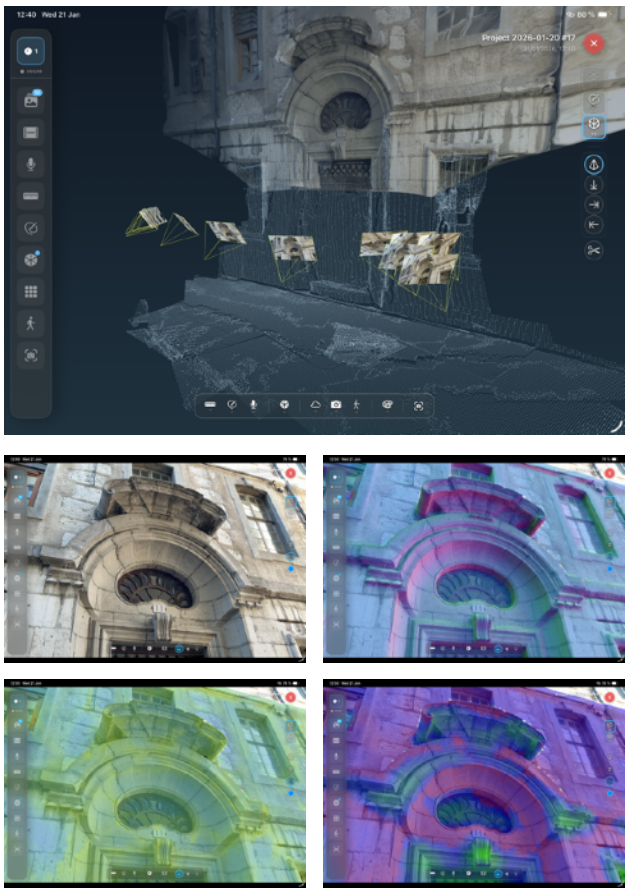

Figure 4. Photogrammetric reconstruction extending beyond LiDAR range and providing high-resolution geometric layers (surface normals, curvature, local density) aligned with the AR reference frame.

In this framework, each captured frame is associated not only with its RGB image, depth map, and confidence map, but also, when available, with a dense photogrammetric mesh registered in the same AR world reference.
Unprojection of 2D annotations is then performed by casting viewing rays from the calibrated camera model and intersecting them with the reconstructed surface using a ray–triangle intersection scheme based on (Möller and Trumbore, 1997). Back-face culling and nearest-hit selection ensure that the resulting 3D anchors correspond to physically visible surface points, yielding metrically accurate positions even in areas beyond LiDAR coverage.
This multi-modal strategy establishes a hierarchical geometric support for semantic anchoring: LiDAR depth is used preferentially for near-range, real-time interaction; when unavailable or unreliable, ARKit's planar and feature-based geometry provides a coarse fallback; and when photogrammetric reconstruction is available, it becomes the

reference for high-resolution and long-range spatial anchoring. Through this progressive refinement, semantic regions drawn directly by the user on image frames can always be lifted into three-dimensional space using the best available geometric support, while remaining embedded in a single, consistent AR reference frame.

To ensure coherent bidirectional projection, the coordinate system of the photogrammetric model is rigidly aligned with the AR world frame using correspondences between camera poses estimated by ARKit and those recovered during reconstruction. A RANSAC-based estimation yields a sub-metric rigid transformation, validated through positional and angular residuals as well as global consistency checks. This alignment allows annotations, measurements, and derived descriptors to circulate seamlessly between real-time AR views, dense 3D geometry, and analytical orthographic representations.

Beyond spatial anchoring, the photogrammetric model also provides access to derived geometric layers such as surface normals, curvature, and local density, which can be visualised and analysed as additional channels of interpretation (Fig. 4).

6. Pillar 3: Interpretation-driven acquisition

6.1. From situated observation to replayable spatio-temporal reasoning

Scientific interpretation in the field does not arise from isolated features alone, but from an exploratory process unfolding in space and time: movement through the site, successive viewpoints, progressive comparison, and the correlation of observations made under specific perceptual and contextual conditions. Meaning thus emerges from a situated trajectory as much as from individual annotation acts.

The third pillar therefore proposes an acquisition paradigm in which data capture is driven by interpretation itself. Rather than aiming at exhaustive geometric coverage, *dür.air* supports deliberate, intention-based recording: each acquisition action results from an explicit observational decision and is immediately associated with semantic content. Every recorded element is thus anchored in the spatial and temporal context in which it becomes scientifically meaningful.

Beyond individual annotations, the system preserves the dynamics of interpretation by modelling each field session as a structured spatio-temporal sequence that combines camera motion, user actions, and semantic events. Sessions are not reduced to collections of images or measurements, but represented as continuous, spatially anchored trajectories along which interpretative acts are situated. In this way, *dür.air* captures not only the outcomes of interpretation, but also the temporal order, spatial context, and exploratory paths through which scientific reasoning is progressively constructed.

6.2. Interpretation-guided acquisition as a spatio-temporal process

From a methodological standpoint, this pillar addresses a key limitation of many 3D documentation workflows: the dissociation between geometric capture and the procedural and temporal dynamics of scientific reasoning. In field sciences, hypotheses and correlations emerge through the succession of viewpoints, movements, and comparisons along an exploratory path; preserving this dynamic is therefore essential for maintaining epistemic continuity.

Dür.air records each session as a structured spatio-temporal sequence combining camera motion, user actions, and semantic events. Acquisition is thus conceived as an event-based process in which interpretative acts are explicitly situated within both the geometry of exploration and the temporal unfolding of the session, enabling the reconstruction and replay of the reasoning process itself.

6.3. Event-based acquisition and observation trajectories

Camera motion is sampled as a continuous six-degree-of-freedom trajectory, while semantically typed events (such as photographs, videorecordings, annotations, measurements, region delineations, or voice comments) are time-stamped and associated with precise positions along this path. Waypoints store camera pose and tracking quality, and bidirectional links connect each event to its spatial and temporal context.

This representation allows raw capture actions to be transformed into semantically qualified observation events (for instance, a photograph becomes an annotation once a region is validated, or a measurement once a geometric operation is completed), while preserving their original ordering and localisation. The resulting structure can be replayed, filtered, and compared across sessions, providing access not only to documented objects, but also to the sequence of viewpoints and interpretative decisions through which knowledge was constructed.

Figure 5 illustrates this principle by showing a field session represented as a three-dimensional camera trajectory together with its synchronised event timeline, where annotations and measurements are explicitly embedded in both space and time.

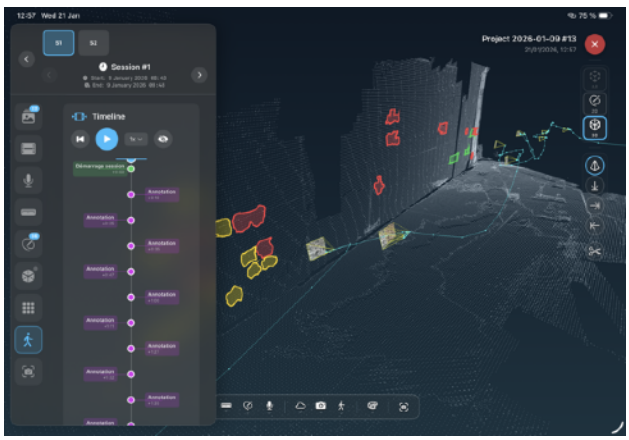

Figure 5. Real-time recording of a field session as a 3D camera trajectory (right) synchronised with an event timeline (left), situating annotation and measurement events along the path of observation.

6.4. Frugality through alignment with scientific intent

Frugality in *dür.air* is not achieved by limiting sensing capabilities, but by structurally aligning acquisition with interpretative intent. Because each capture requires deliberate user action and is immediately contextualised semantically, the system favours relevance and interpretability over indiscriminate data accumulation.

By preserving the full spatio-temporal envelope of observation sessions, the platform supports cumulative and longitudinal reasoning. Successive visits do not merely add new datasets, but extend an existing observation structure in which trajectories, events, and annotations can be replayed, confronted, and reinterpreted. Field documentation thus becomes a progressive construction of knowledge, grounded in the continuity of situated observation and in the explicit memorisation of the processes through which interpretations emerge.

7. System-Level Enablers

7.1. Structured and extensible description

In *dür.air*, semantic information is modelled through a template-driven architecture in which annotations are defined by typed, versioned schemas rather than free-form text. Templates specify controlled vocabularies, validation rules, and domain-specific fields, ensuring that semantic enrichment remains consistent across sessions, operators, and projects, while remaining extensible as scientific protocols evolve.
Each annotation is further represented as a multi-scale entity that tightly couples its 2D delineation on images, its 3D spatial anchoring in the AR reference frame, and a set of automatically extracted descriptors (Fig. 6).
This unified data structure links pixel-level evidence to metric geometry and to expert-defined semantics, preserving the full geometric lineage of the observation. Three complementary families of descriptors are associated with each annotated region:

- *geometric descriptors* derived from the 3D anchor points (size, orientation, planarity, compactness, elongation, volumes);
- *visual descriptors* computed on-device from the corresponding image crop (colour statistics, contrast, texture, entropy, dominant chromatic components);
- *semantic descriptors* provided by the user through the template system (typed fields, vocabulary terms, validated values, …).

This integrated model transforms annotations from simple graphical marks into structured, queryable, and analysable entities, in which conceptual interpretation is explicitly grounded in both geometry and appearance and anchored within a persistent spatio-temporal reference system. By formalising annotation as typed, multi-scale semantic objects embedded in data lineages and provenance structures, the approach supports structured querying, richer context management, and analytical exploration of semantic annotation graphs, as explored in recent work on bridging provenance and semantic interpretation in 3D cultural heritage workflows (Guillem et al., 2025).

7.2. Georeferencing and offline-first operation

Projects can be associated with absolute geographic coordinates through on-device GNSS, and standard EXIF metadata (position, time, camera intrinsics) are propagated to exported imagery, ensuring interoperability with GIS environments and multi-scale spatial analyses. The system also supports relative and absolute spatial point sets interpreted as ground control points (GCPs), allowing the AR reference frame and derived models to be embedded within traditional surveying workflows.
All acquisition, alignment, and annotation processes are performed locally on the tablet following an offline-first design, while network-dependent services (e.g. geocoding, external vocabularies, cloud storage) rely on deferred synchronisation. This ensures full autonomy in the field while maintaining compatibility with external platforms and FAIR-oriented archival infrastructures.

8. Discussion and Conclusion

This work addresses a structural limitation of many contemporary field documentation workflows: the dissociation between in situ observation, data acquisition, and interpretation. While advances in photogrammetry and mobile sensing have greatly improved geometric accuracy, they often reinforce sequential pipelines in which interpretation is deferred and progressively detached from the spatial and temporal conditions of observation.

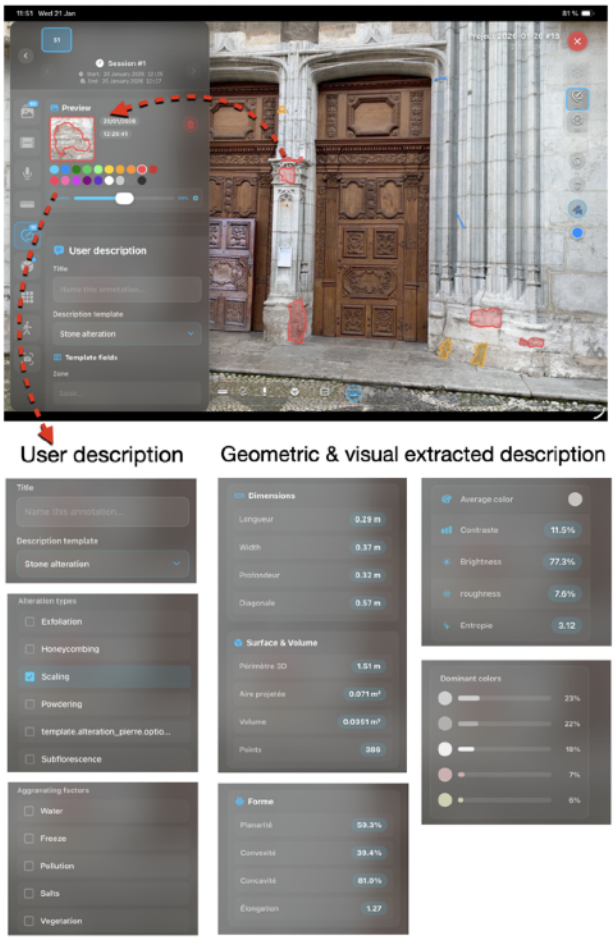

Figure 6. Example of a single spatial annotation combining user-defined semantic fields (left) with automatically extracted geometric and visual descriptors (right), illustrating the multi-scale and multi-modal structure of the annotation model.

The proposed approach re-establishes a tight coupling between these processes by embedding interpretation within acquisition and by preserving spatial and temporal continuity across sessions. Multi-temporality is treated as a first-class methodological dimension: observations persist within a stable spatial reference, enabling cumulative enrichment and diachronic reasoning, which is particularly relevant for disciplines grounded in longitudinal fieldwork such as archaeology, conservation, and landscape studies. The interpretation-driven acquisition strategy further shifts the focus from exhaustive capture to semantic efficiency.
By linking each acquisition act to explicit observational intent, the system reduces uncontextualised data production while increasing the density, traceability, and reusability of scientific meaning, in line with current concerns regarding sustainable and interpretable data practices.
Beyond methodological aspects, *dür.air* is conceived as an evolutive framework, adaptable to current and future embedded sensors and avoiding any irreversible fixation of models, spatial registrations, or projection schemes. All representations remain recalculable and exportable toward established disciplinary and digital practices (e.g. orthomosaics, ground control points, spatialised images, 3D point clouds and models), supporting long-term interoperability and FAIR principles, even when

currently operating on initially closed hardware ecosystems such as mobile Apple devices.
Several limitations remain. Persistent spatial anchoring depends on environmental stability and visual richness, and mobile hardware still constrains real-time processing and autonomy. Moreover, the workflow presupposes an active interpretative engagement, which may require methodological adaptation and training. Future developments will address multi-user and collaborative scenarios, deeper integration with formal semantic models and knowledge graphs, and the use of machine learning techniques to assist annotation and similarity-based recognition.
In conclusion, this work advocates a shift in the design of digital tools for field sciences: from systems primarily aimed at producing ever more complete geometric replicas, toward environments that explicitly integrate observation, interpretation, and long-term spatio-temporal memory as core components of scientific reasoning.

Acknowledgements

This research was primarily supported by the European Research Council through the ERC-2021-Advanced Grant project nDame_Heritage (Grant Agreement No. 101055423). The work also benefited from support from the French National Research Agency (ANR) under the Investissements d'Avenir programme integrated into France 2030, through the Foundation for Cultural Heritage Sciences (ANR-17-EURE-0021) and the Equipex+ ESPADON digital facility (ANR-21-ESRE-0050). The development of *dür.iar* is furthermore aligned with the objectives of the European Research Infrastructure for Heritage Science (E-RIHS), in particular its DIGILAB platform.

References

Abergel, V ., Manuel, A., Pamart, A., Cao, I., De Luca, L., 2023: Aïoli: A reality-based 3D annotation cloud platform for the collaborative documentation of cultural heritage artefacts. Digital Applications in Archaeology and Cultural Heritage, 30.

Abergel, V., Jacquot, K., De Luca, L., & Veron, P., 2019: Towards a SLAM-based augmented reality application for the 3D annotation of rock art. Int. Journal of Virtual Reality, 19(2).

Antón, D., Mayoral-Valsera, J., Simón-Vallejo, M.D., Parrilla-Giráldez, R., Cortés-Sánchez, M, 2025: Built-in smartphone LiDAR for archaeological and speleological research, Journal of Archaeological Science, Vol. 181.

Bekele, M.K.,, Pierdicca, R., Frontoni, E.,, Malinverni, E., Gain, J., 2018: A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage. J. Comput. Cult. Herit. 11, 2.

Brandi, C., 1963: Teoria del restauro. Edizioni di Storia e Letteratura, Roma.

Corner, J., 1999: The agency of mapping: speculation, critique and invention. In: Mappings. Reaktion Books, London, 213–252.

Courty, M.A., Goldberg, P., Macphail, R., 1989: Soils and Micromorphology in Archaeology. Cambridge University Press, Cambridge.

Dell'Unto, N., Landeschi, G., Apel, J., Poggi, G., 2017: 4D recording at the trowel's edge: using three-dimensional simulation platforms to support field interpretation. Journal of Archaeological Science: Reports 12, 632-645

Guillem, A., Abergel, V., Roussel, R., Comte, F., Pamart, A., De Luca, L., 2025: Bridging the Provenance Knowledge Gap Between 3D Digitization and Semantic Interpretation. Heritage, 8(11), 476.

Guo, A., Canberk, I., Murphy, H. Monroy-Hernández, A., Vaish, R., 2019: Collaborative and Persistent Augmented Reality Experiences. Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. Vol. 3

Hodder, I., 1999: The Archaeological Process: An Introduction. Blackwell, Oxford.

Ingold, T., 2011: The Perception of the Environment: Essays on Livelihood, Dwelling and Skill. Routledge, London.

Kędziorski, P., Jagoda, M., Tysiąc, P., & Katzer, J., 2024: An Example of Using Low-Cost LiDAR Technology for 3D Modeling and Assessment of Degradation of Heritage Structures and Buildings. Materials, 17(22), 5445.

Klein, G., Murray, D., 2007: Parallel Tracking and Mapping for Small AR Workspaces. IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan.

Letellier, R., Schmid, W., LeBlanc, F., 2007: Recording, Documentation and Information Management for the Conservation of Heritage Places: Guiding Principles. Getty Conservation Institute, Los Angeles.

Lucas, G., 2012: Understanding the Archaeological Record. Cambridge University Press, Cambridge.

Manuel, A. De Luca, L., Veron, P., 2014: A Hybrid Approach for the Semantic Annotation of Spatially Oriented Images. Int. Journal of Heritage in the Digital Era, 3 (2).

Möller, T., Trumbore, B., 199: Fast, Minimum Storage Ray-Triangle Intersection. Journal of Graphics Tools, 2(1).

Pamart, A., Morlet, F., De Luca, L., Veron P., 2020: A robust and versatile pipeline for automatic photogrammetric-based registration of multimodal cultural heritage documentation. Remote Sensing 12 (12), 2051

Ponchio, F., Callieri, M., Dellepiane, M., Scopigno, R., 2020: Effective Annotations Over 3D Models, Computer Graphics Forum. 39 (2020) 89–105.

Remondino, F., Campana, S. (eds.), 2014: 3D Recording and Modelling in Archaeology and Cultural Heritage. BAR International Series 2598, Archaeopress, Oxford.

Sapirstein, P. (2016). Accurate measurement with photogrammetry at large sites. Journal of Archaeological Science, 66.

Teppati Losè, L., Spreafico, A., Chiabrando, F., Giulio Tonolo, F., 2022: Apple LiDAR Sensor for 3D Surveying: Tests and Results in the Cultural Heritage Domain. Remote Sensing, 14(17).