

GeoSynTwin: Knowledge-Graph-Constrained Synthetic Data Generation for Multidimensional Geospatial Digital Twins

Akram Hakiri¹[0000-0001-7151-5499] ^{*}, Vincent Lalanne¹[0000-0002-7907-527X],
Manuel Munier¹[0000-0003-0282-4508], Richard Chbeir¹[0000-0003-4112-1426],
and Laurent Gallon¹[0000-0003-4252-1421]

University of Pau and Pays de l'Adour, E2S, UPPA, LIUPPA, College STEE, Anglet
`firstname(dot)lastname(at)univ-pau.fr`

Abstract. Geospatial digital twins support environmental analysis by integrating heterogeneous spatial and temporal data. However, their reliability is limited by real-world datasets that are often sparse, spatially biased, predominantly static, and lacking coverage of rare events. Although synthetic data generation can mitigate these gaps, existing methods rarely guarantee physical, topological, and ecological consistency or provide traceable data provenance. We present GeoSynTwin, a knowledge-graph-driven framework for generating constraint-compliant, provenance-aware synthetic data for geospatial digital twins. GeoSynTwin leverages a geospatial knowledge graph aligned with GeoSPARQL, SOSA, PROV-O, and Darwin Core to supply scenario-specific context and encode five classes of constraints: geometric, physical, semantic, temporal, and provenance. These constraints govern four specialized generators via differentiable penalties in a dynamic GAN, grammar-constrained decoding, vision-language model validation and repair, and hard transition masks in a convolutional scene model. Across seven MultiGIS modalities—including flood-depth sequences, Sustainable Drainage Systems (SuDS) layouts, species trajectories, sensor streams with labeled artifacts, and geo-referenced radiance fields, the framework publishes each synthetic product as a SpatioTemporal Asset Catalog (STAC) item with PROV-O lineage and explicit synthetic labeling before streaming it into an Eclipse Ditto digital twin. We evaluate GeoSynTwin across two CHIST-ERA AI4MultiGIS pilot studies: urban flooding and SuDS in Chelmsford, UK, and invasive crayfish spread in Romania. By coupling generation with explicit domain constraints and end-to-end lineage, GeoSynTwin establishes a unified, traceable pipeline for multimodal geospatial data synthesis.

Keywords: Synthetic data · Digital twin · Knowledge graph · GeoAI · MultiGIS · Generative models.

^{*} Corresponding author.