URBES an Intelligent Geospatial Platform for the Identification and Valorization of Critical Raw Materials

Carlo Cipolloni[1[0000-0001-7652-9814]], Giada Possemato[1[0009-0006-4242-3898]], Ignazio Scano [1[0009-0005-0047-7620]], Giulio Gatti[2], Tiziano Olivieri[2], Paolo Gigliarelli Fiumi[2]

[1]Istituto Superiore per la Protezione e la Ricerca Ambientale, Roma, Italy
[2]SAS Institute, Milano, Italy
carlo.cipolloni@isprambiente.it

Abstract.
The URBES platform addresses data fragmentation to support Critical Raw Material recovery through Urban Mining. Built on a 5-layer AI-GIS architecture, it integrates Natural Language Processing, U-Net cartographic segmentation, and closed-domain Retrieval-Augmented Generation with strict spatial validation. The system transforms unstructured historical archives into FAIR geodatabases, offering a robust decision support tool.

Parole chiave: AI, Critical Raw Materiale, RAG, Architecture.

1 Introduction and Strategic Context

The secure, sustainable, and ethically responsible supply of Critical Raw Materials (CRMs) and Strategic Raw Materials (SRMs) represents a fundamental prerequisite for ensuring European technological sovereignty, industrial competitiveness, and the successful implementation of the green transition. Within this framework, both national and European policies aimed at strengthening supply chain resilience increasingly recognize the importance of **Urban Mining**, defined as the recovery of resources from end-of-life products and urban infrastructures, as well as the valorization of historical extractive waste deposits [1].

These secondary sources constitute a potentially strategic reservoir of raw materials; however, their effective exploitation is hindered by the fragmentation, heterogeneity, and limited accessibility of the available information. A significant proportion of the knowledge related to abandoned mining sites, extractive waste deposits, and recoverable resources is currently preserved within historical documentation, technical archives, and paper-based reports that are only marginally interoperable with modern information systems.

To address this challenge, the URBES (Urban Mining and Extractive Waste Information System) project was funded through the Italian National Recovery and Resilience Plan (PNRR), Mission 7 RePower program. The initiative is structured into four coordinated work packages. In particular, Work Package 2, entitled *Intelligent Knowledge Platform and WebGIS*, focuses on the design and implementation of a

geospatial cognitive infrastructure capable of integrating geological, territorial, and documentary data while simultaneously enhancing the value of historical information assets that are often dispersed or unstructured.

The primary objective of the URBES platform is to overcome barriers to information accessibility and usability through the development of an integrated, queryable, and continuously updatable geospatial knowledge base. To achieve this goal, a multilayer architecture comprising five functional levels has been developed. These layers are responsible for information acquisition and management, automated textual content analysis through Natural Language Processing (NLP) techniques, statistical and spatial data validation, and the implementation of a cognitive engine based on Retrieval-Augmented Generation (RAG) approaches [2] and Large Language Models (LLMs).

The entire ecosystem is coordinated by a multi-agent orchestration framework that enables users to query documents, geodatabases, and geographic information through natural language interactions within a unified interface. This architecture promotes interoperability among heterogeneous information sources and supports the generation of traceable and scientifically verifiable responses through dedicated data validation and quality-control procedures, thereby reducing the risk of interpretative errors commonly associated with generative Artificial Intelligence systems.

1.1 The Challenge of Geological and Mining Data Fragmentation

The Italian geological and mining information heritage has historically been characterized by significant fragmentation, both from structural and technological perspectives. A substantial portion of information concerning historical production activities, accumulated volumes, and ore grades is contained within so-called non-conventional datasets, including historical paper documentation, mining permits and leases records, technical exploration reports, mine survey maps, and geological bulletins preserved over more than a century within national and regional archives.

Despite their considerable scientific and informational value, these documentary assets generally lack both semantic and geographic indexing, making their identification, integration, and analysis through conventional information technologies particularly challenging at a national scale. Table 1 summarizes the main operational challenges associated with the management and processing of the various categories of raw data currently preserved and managed by the Geological Survey of Italy (ISPRA).

To overcome these limitations, the URBES methodology defines a high-fidelity digital transformation workflow aimed at converting unstructured documents and raster cartography into structured, georeferenced, topologically validated datasets that can be directly interpreted by machine-based systems. The approach incorporates and operationalizes the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [6], with particular emphasis on interoperability, which is regarded as a key enabling factor for the integration of heterogeneous information sources.

Within this framework, the automated extraction of toponyms, mining sites, and geological-mining entities through Natural Language Processing (NLP) techniques is coupled with dynamic geocoding procedures capable of transforming unstructured textual information into explicit geographic entities. As a result, isolated descriptive

references are converted into queryable spatial primitives that can be exploited within WebGIS environments and advanced territorial analysis systems.

Table 1. Data source types.

Raw Data Type	Key Limitations of Conventional Processing Workflows
More than 2,500 PDF documents (30-150 MB each; historical scans, images, and native PDF files)	Limited indexing of scanned content; difficulties in preserving the structural relationships among text, tables, figures, and appendices; absence of explicit geographic metadata associated with place names, mining production records, and extractive entities.
More than 200 raster geological maps (historical and thematic cartography not available in vector format)	Presence of graphical and symbolic layers lacking native georeferencing; inability to perform advanced spatial queries and topological analyses; limited computational usability of geological legends and mapped geological units.
JSON-native geodatabases and SQL tables (heterogeneous and decoupled management databases)	Misalignment among the conceptual and syntactic schemas adopted by regional and national registries; difficulties in integrating structured datasets with qualitative and extractive information derived from historical documentation.

2 The URBES platform architecture

During the preliminary phases of the URBES project, the technological infrastructure was initially conceived as a three-layer architectural framework comprising a data ingestion layer, a data analytics layer, and a Retrieval-Augmented Generation (RAG) layer exposed to end users. However, subsequent analyses of computational requirements and the scientific validation needs defined by ISPRA led to the development of a more articulated solution based on five distinct functional layers (Figure 1).

The main architectural evolution involved the separation between the Artificial Intelligence-driven integration and analytics layer and the layer dedicated to geostatistical analysis and spatial validation. This design choice was motivated by the need to isolate heterogeneous computational workloads characterized by different hardware and performance requirements, while introducing a rigorous geospatial data quality-control mechanism based on the Human-in-the-Loop (*HITL*) paradigm.

In particular, document processing, computer vision tasks, and semantic embedding generation require highly parallelized GPU-accelerated infrastructures, whereas geostatistical analyses, GIS processing workflows, and topological verification procedures rely on CPU-oriented architectures with intensive access to spatial databases. Furthermore, the separation of these two domains implements a dedicated

"anti-hallucination barrier," whereby every geographic entity generated or inferred by Artificial Intelligence models undergoes deterministic validation procedures before being incorporated into the system's knowledge base.



Fig. 1. Five-layer functional architecture of the URBES AI-GIS platform.

The architecture implements a progressive data transformation process that converts heterogeneous and unstructured information into formally validated geospatial knowledge. The first layer is dedicated to the acquisition, normalization, and preparation of information sources originating from ISPRA archives and associated documentary repositories. The infrastructure currently manages more than 2,500 PDF documents, over 200 raster geological maps, and numerous structured datasets stored in relational databases and JSON-native geodatabases. At this stage, advanced Optical Character Recognition (OCR) processes and document layout analysis models are applied to identify the logical structure of document pages and preserve the relationships among textual content, tables, legends, and annotations. Simultaneously, data normalization procedures, scan correction, noise reduction, and document integrity verification activities are performed.

The second layer implements semantic extraction, vectorization, and automated image analysis functionalities through high-performance infrastructures equipped with state-of-the-art GPU accelerators (e.g., NVIDIA H100). Textual contents are processed through NLP pipelines based on overlapping chunking strategies and state-of-the-art embedding models (BGE and E5) [5], enabling the extraction of domain-specific entities such as mining-related toponyms, mineral categories, and quantitative production parameters. The resulting semantic vectors are stored and indexed using PostgreSQL with the pgVector extension, enabling similarity-based semantic retrieval operations. In parallel, a U-Net convolutional neural network architecture [3], specifically trained for cartographic segmentation tasks, performs the automated vectorization of historical geological maps, transforming raster elements into

georeferenced geographic primitives while converting symbols and legends into queryable metadata. This layer also integrates FAIR-compliant geocoding services to ensure data interoperability and reusability.

The third layer represents the scientific validation and geospatial quality-assurance core of the entire ecosystem. Through the integration of SAS Viya with SAS Intelligent Decisioning, used for middleware integration system, SAS RAM Retrieval Agent Manager, used for LLM queries, PDF ingestion and queries to POSTGRES using MCP TOOLS, PostGIS, and QGIS platforms, advanced statistical analyses, kriging-based geostatistical estimations, and three-dimensional block-modeling procedures are performed to quantify critical and strategic mineral resources contained within extractive waste deposits. Concurrently, a comprehensive suite of topological validation procedures verifies geometric consistency, the integrity of cartographic reference systems, the absence of polygon anomalies, and the administrative coherence of geographic locations. Only datasets that successfully pass all validation procedures are promoted to the subsequent layer and made available as certified data collections. The adoption of Open Geospatial Consortium (OGC) standards, including WMS, WFS, and OGC API Features, further ensures full interoperability with external geospatial infrastructures.

The fourth layer constitutes the system's certified cognitive repository and provides the foundation for the Retrieval-Augmented Generation (RAG) engine [2]. Optimized for GPU-based inference, the infrastructure integrates the open-weights Llama Gemma4-31B language model with a knowledge base composed of vector collections, relational geodatabases, and domain-specific operational rules pertaining to the mining sector. The adoption of a closed-domain approach restricts the model's reasoning capabilities exclusively to information validated within the geo-mining ecosystem, significantly reducing the risk of hallucinations while ensuring high levels of scientific reliability.

The fifth layer performs intelligent orchestration functions and serves as the primary interaction point for end users. A central component, referred to as the Orchestrator Agent, analyzes user intent, coordinates the parallel execution of specialized agents, and integrates outputs generated by the various subsystems. Retrieval settings have been created to define the logic for interpreting requests and the response procedures, including routing systems specialized in directing the analysis only to the collections dedicated to the requested topic. Access to the ecosystem functionalities is provided through a GTT-based (Generative Trained Transformer like GPT) virtual assistant integrated within a multilingual conversational WebGIS platform. This environment enables natural language queries and delivers, through a unified interface, textual responses, interactive cartographic visualizations, and direct links to the original documentary sources. Within this configuration, the URBES AI platform implements a workflow based on the convergence of two complementary processes: the Semantic Retrieval Path and the Geometric Validation Path, ensuring that every generated response is simultaneously semantically relevant and geographically validated.

3 Considerations and Future Developments of the Virtual Agent

The synergistic integration of advanced Computer Vision techniques (U-Net), Layout Analysis technologies (Tesseract and PaddleOCR) [4], and closed-domain Retrieval-Augmented Generation (RAG) [3] systems based on OLlama-Gemma4-31B has enabled ISPRA to transform more than a century of historical documentation, previously unstructured and difficult to access, into a modern, dynamic, and queryable geographic information infrastructure. This process of digitization and knowledge enhancement generates substantial operational and strategic benefits.

First, it supports strategic autonomy and circular economy principles through the systematic mapping of mining waste deposits, enabling the identification of sites with high potential for the recovery of Critical Raw Materials (CRMs) and Strategic Raw Materials (SRMs), thereby reducing dependence on foreign supply sources. At the same time, the system dramatically decreases the time required for archival consultation and information georeferencing, transforming activities that previously demanded weeks of manual processing into real-time query operations. Furthermore, the platform provides robust support for environmental policies and territorial governance by delivering an integrated knowledge framework that combines geochemical data, information on mining sites, and urban landfill inventories, thereby enhancing remediation, monitoring, and environmental risk mitigation activities.

The most significant methodological innovation lies in the ability to combine the inferential power of Large Language Models (LLMs) with the rigor, traceability, and determinism traditionally required by Geographic Information Systems (GIS).

The scientific robustness of the architecture is founded upon three fundamental principles. The first is the concept of vector-geometric informational duality, whereby every semantic entity identified by the language models is uniquely associated with its corresponding spatial geometry within the PostGIS geodatabase, ensuring a direct linkage between textual knowledge and geographic representation. The second principle is the systematic elimination of spatial hallucinations through a dedicated validation layer that subjects every geographically referenced output generated by Artificial Intelligence models to rigorous topological verification procedures, including the validation of coordinate reference systems, geometric integrity, and administrative consistency, before integration into the knowledge base. The third principle concerns the adoption of an open and interoperable ecosystem, compliant with OGC standards (WMS, WFS, and OGC API Features) and complemented by OpenAPI endpoints, thereby ensuring full adherence to FAIR principles and enabling data access and reuse by other institutional WebGIS systems.

The modular structure of the AI Platform, organized according to a multilayer architecture, also demonstrates a high degree of scalability and methodological transferability. The framework developed within the URBES project for Urban Mining and the valorization of Critical Raw Materials can be regarded as a replicable operational paradigm applicable to numerous domains of applied geosciences, including natural hazard assessment, groundwater resource management, spatial planning, and the protection of critical national infrastructures.

From this perspective, the introduction of the Geological AI Framework represents a significant advancement in the Institution's capability to leverage and exploit its geospatial information assets, transforming them into a strategic instrument for decision support and for the implementation of national policies oriented toward sustainability, environmental security, and the efficient management of natural mineral resources.

Riferimenti bibliografici

1. Blengini, G. A., Mathieux, F., Mancini, L., & European Commission, Joint Research Centre. (2019). Recovery of critical and other raw materials from mining waste and landfills: State of play on existing practices key challenges and opportunities. Publications Office of the European Union. EUR 29868 EN.
2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems (NeurIPS 2020), Vol. 33, pp. 9459–9474.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Lecture Notes in Computer Science, Vol. 9351, pp. 234–241. Springer, Cham.
4. Smith, R. (2007). An overview of the Tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Vol. 2, pp. 629–633. IEEE.
5. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv preprint arXiv:2212.03533
6. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018.