EUROPEAN BIG DATA VALUE FORUM

Karel Charvat

Fusing Language Models and Earth Observation for the Next-Gen AKIS

ORGANISED BY

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH

European Commission

DTU

SDU University of Southern Denmark

ebdvf.eu

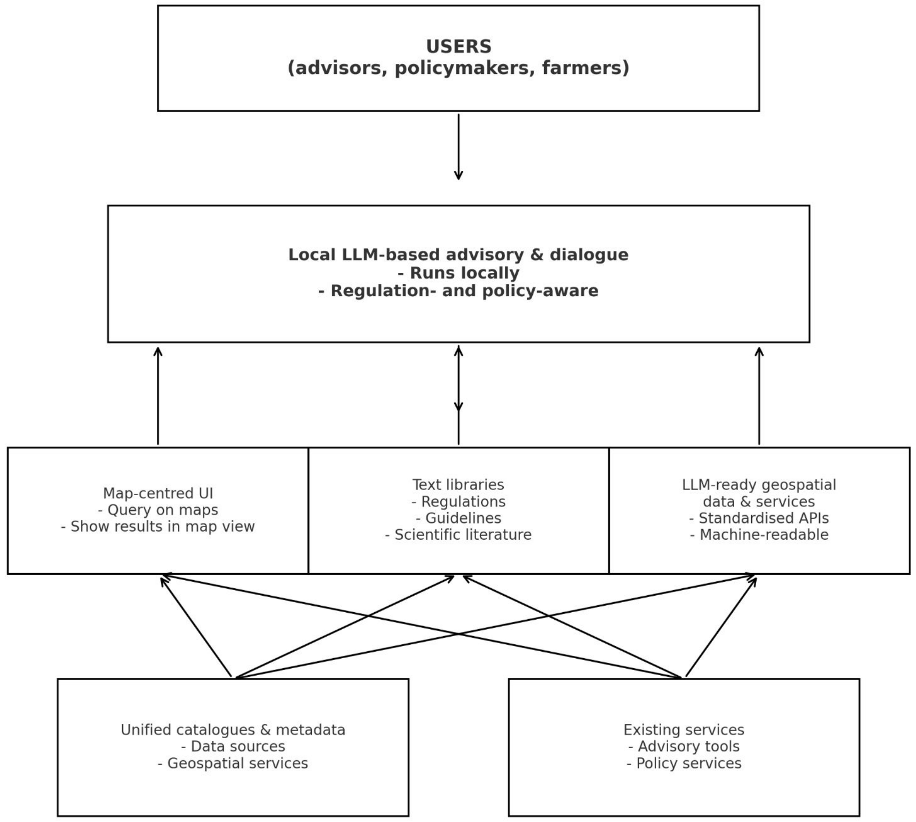
#EBDVF25

From Data to Wisdom: JackDaw in Next-Generation AKIS

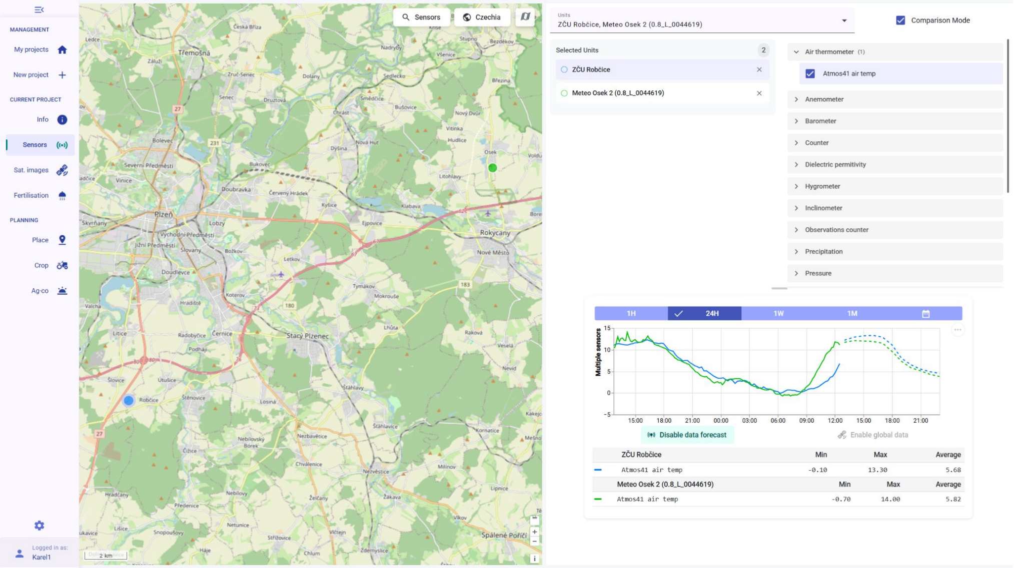
- Traditional AKIS focuses on data aggregation, with limited contextual reasoning and multilingual interaction.
- New LLMs enable cross-lingual, explainable advisory dialogue grounded in domain knowledge.
- JackDaw integrates LLMs with GIS, IoT sensors and Sentinel data for location-specific recommendations.
- Thematic RAGs connect advisory logic to scientific evidence, legislation, regulations, geo-data and policy frameworks.
- Result: regulation-aware, spatially contextual decision support aligned with evolving CAP and Green Deal requirements.

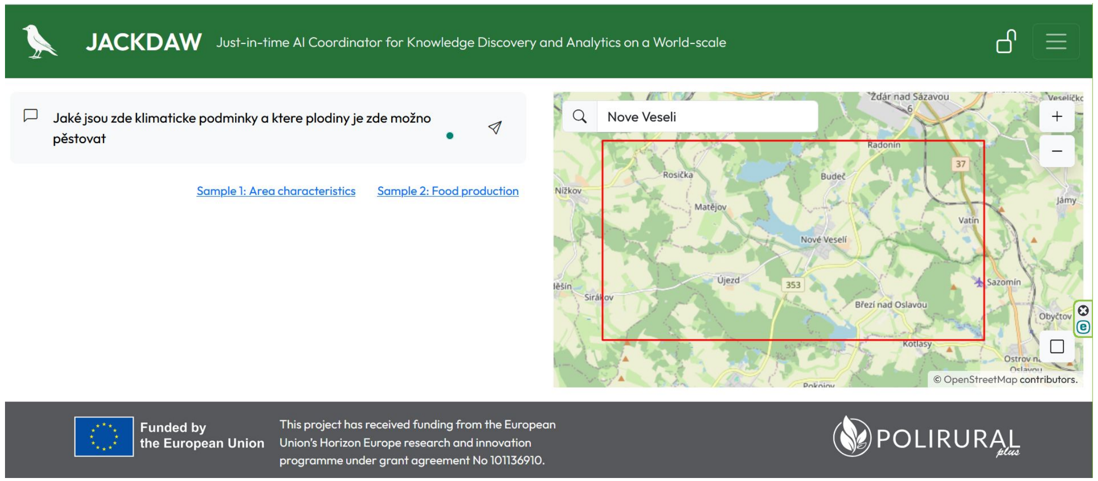
Current work and key technological challenges

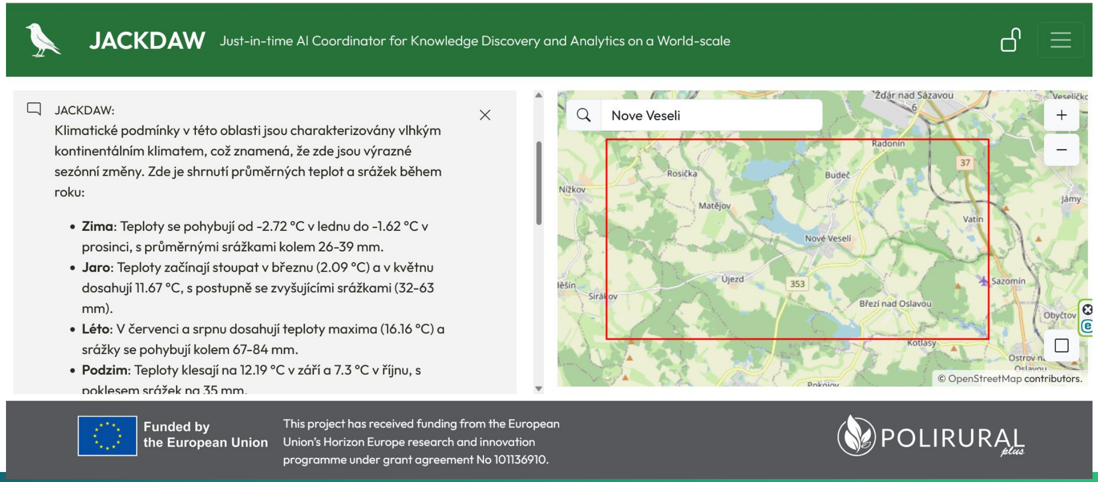
- Adapt existing advisory and policy services for LLM-based interaction.
- Curate domain-specific text libraries (regulations, guidelines, scientific literature).
- Expose geospatial data and services in machine-interpretable formats for LLM tools.
- Design map-centred interfaces for spatial querying and result visualisation.
- Build unified catalogues and metadata for available data and services.
- Deploy and optimise locally running LLMs to ensure sovereignty and compliance.



Adapt existing advisory and policy services for LLM-based interaction

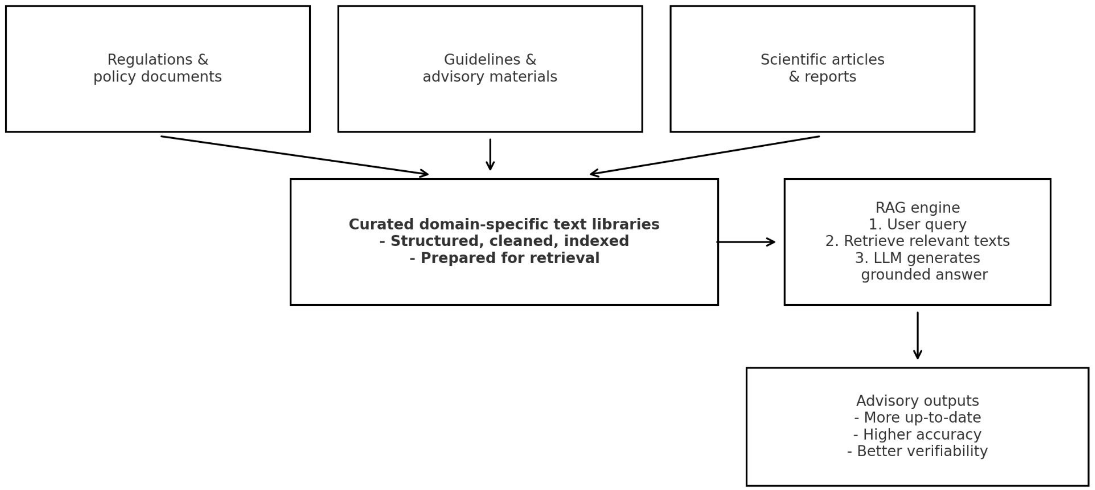






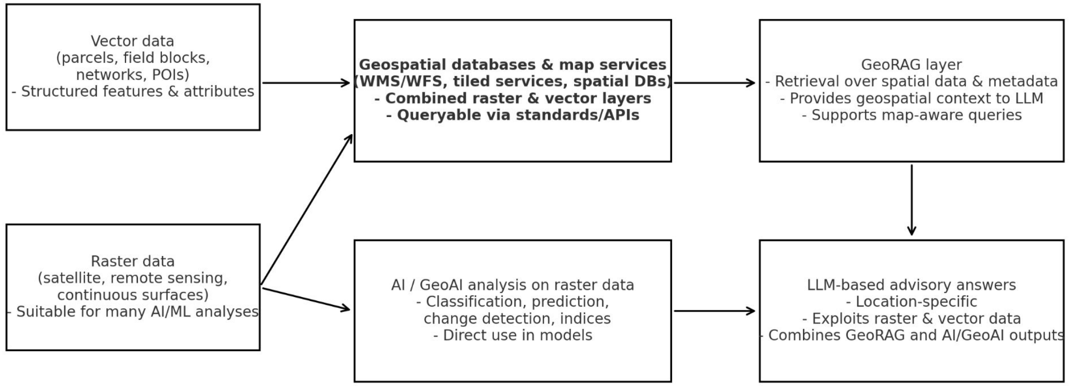
Curate domain-specific text libraries

- Around 2007, the idea emerged to connect quantitative data analysis with automated processing of prescriptions, regulations, directives and scientific articles, but it remained largely a technological utopia.
- Today, Retrieval-Augmented Generation (RAG) enables this by combining language models with targeted search in external document collections and databases.
- The model first retrieves relevant domain documents (regulations, guidelines, scientific papers) and then composes answers grounded in these sources.
- This approach increases the timeliness, factual accuracy and verifiability of advisory outputs.



Geospatial data and services

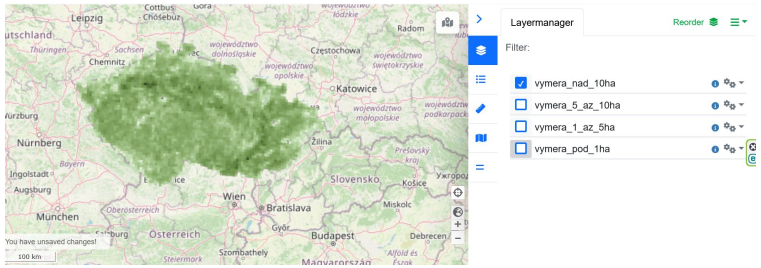
- Geospatial information systems work with both raster and vector data (satellite imagery, remote sensing products, cadastral parcels, field blocks, infrastructure networks).
- Raster data are particularly suitable for many AI and ML analyses (e.g. classification, change detection, yield prediction).
- To fully exploit language models, geospatial data must be accessible in a structured, queryable form (features, attributes, time, uncertainty, provenance).
- GeoRAG solutions link LLMs with spatial databases and map services, allowing the model to retrieve and reason over relevant geospatial layers and services when generating answers.

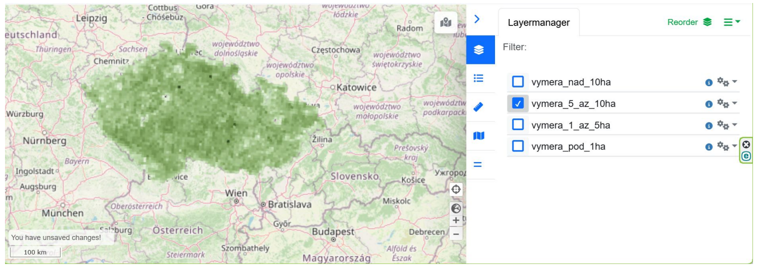


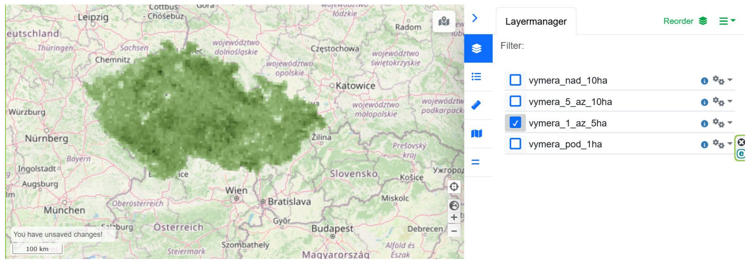
Centrality measure

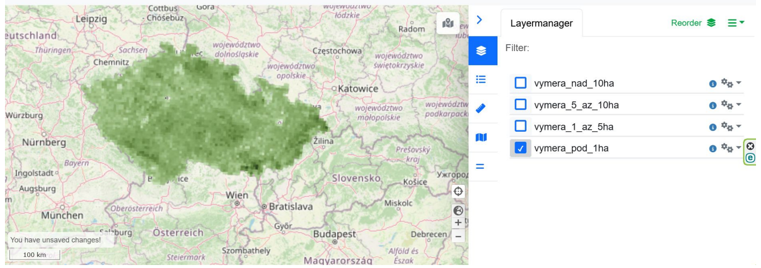
Centrality is a spatial indicator that quantifies the level of presence of a selected feature (e.g. roads, services, infrastructure, specific land use) in the surroundings of a given point or grid cell.

- For each point, the measure aggregates information about features in a defined neighbourhood (e.g. within a radius or along a network) and expresses their density, intensity or accessibility.
- In rasterisation workflows, centrality measures allow vector features to be transformed into continuous raster surfaces that capture how strongly a feature is represented around each location.

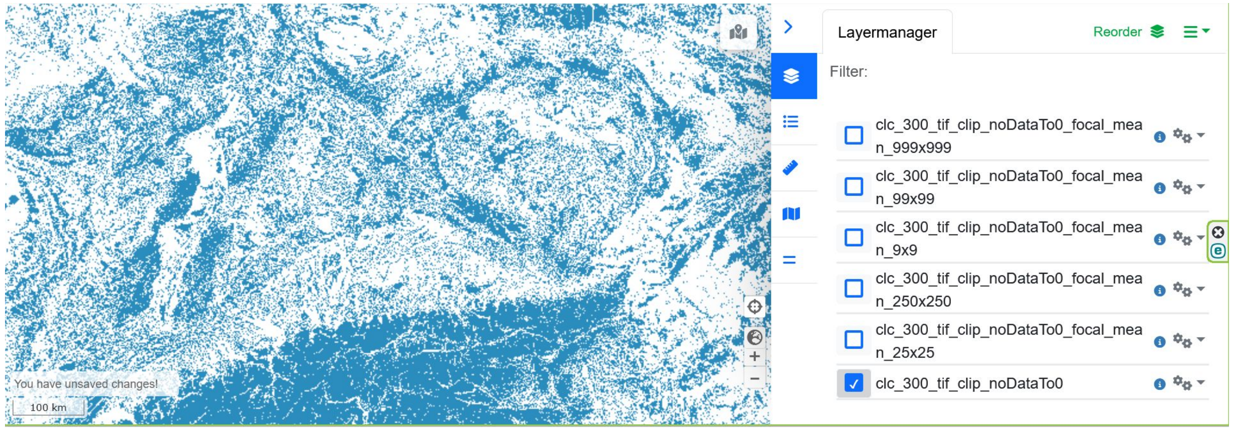


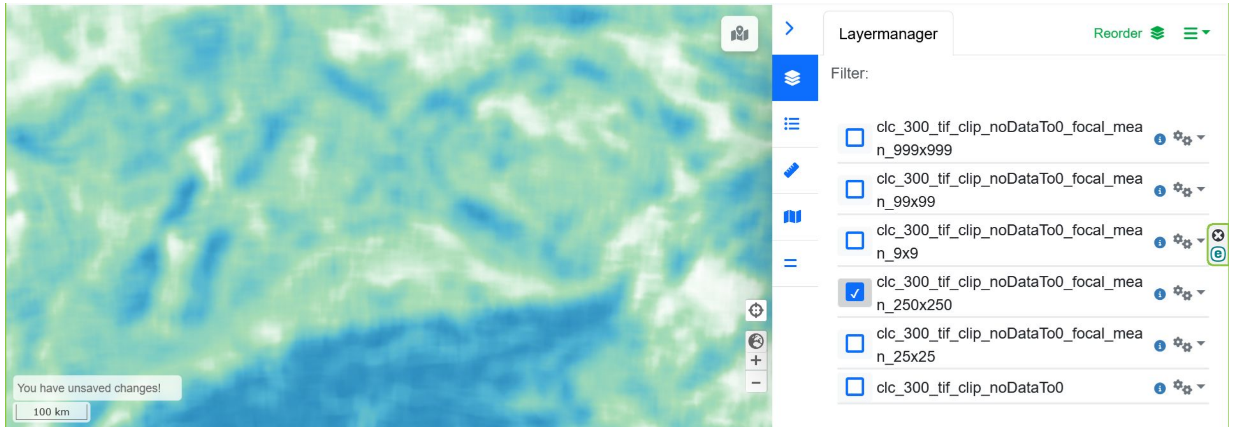


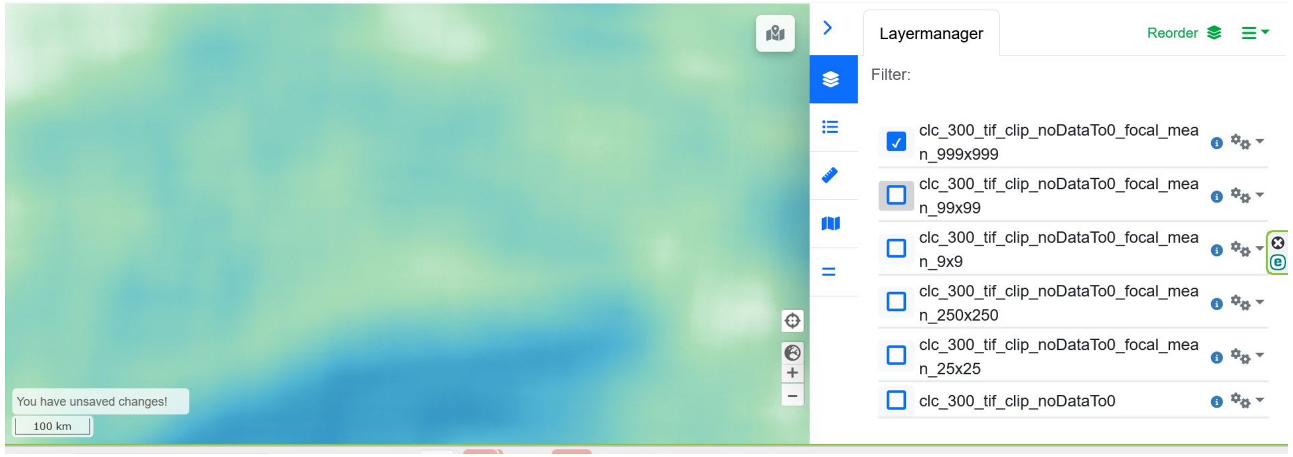




Centrality measure can be in different scales

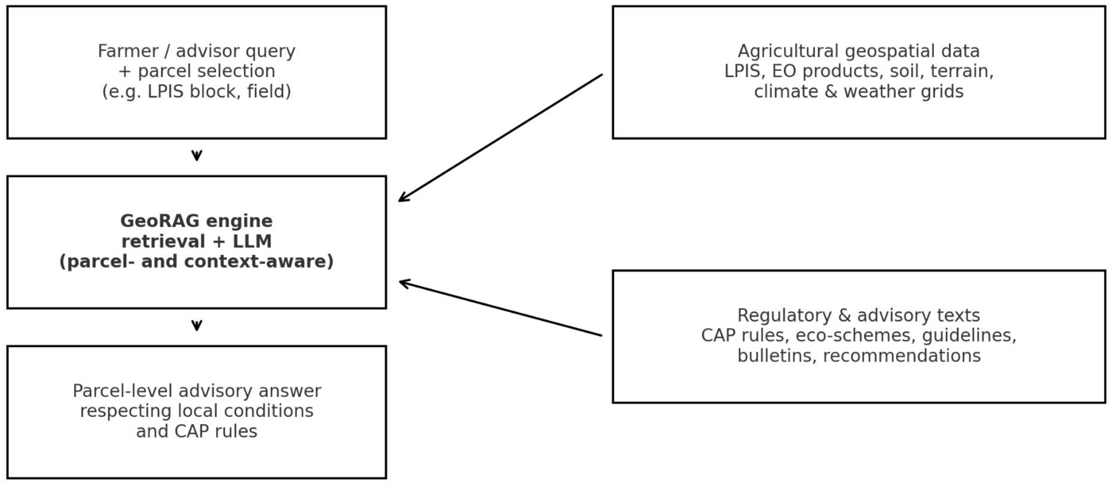






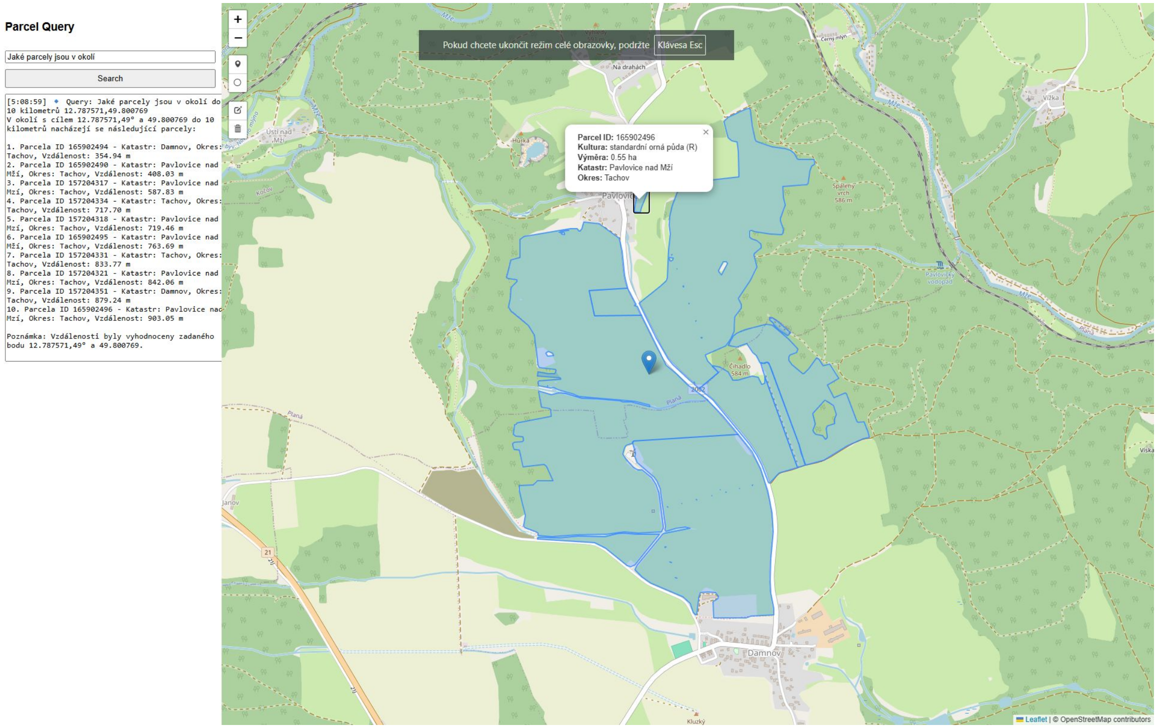
GeoRAG

- GeoRAG combines a large language model with retrieval from agricultural geospatial data (e.g. LPIS/land parcel data, Earth Observation, soil maps, climate and meteorological grids).
- Before generating an answer, the model first searches spatial databases, metadata records and textual documents (e.g. CAP rules, guidelines, advisory bulletins) for the specific location and crop.
- The response is then generated on top of these retrieved geospatial and textual sources, so that recommendations respect parcel boundaries, local conditions and regulatory constraints.
- This enables parcel-level, context-aware advisory dialogue instead of generic, location-agnostic answers.



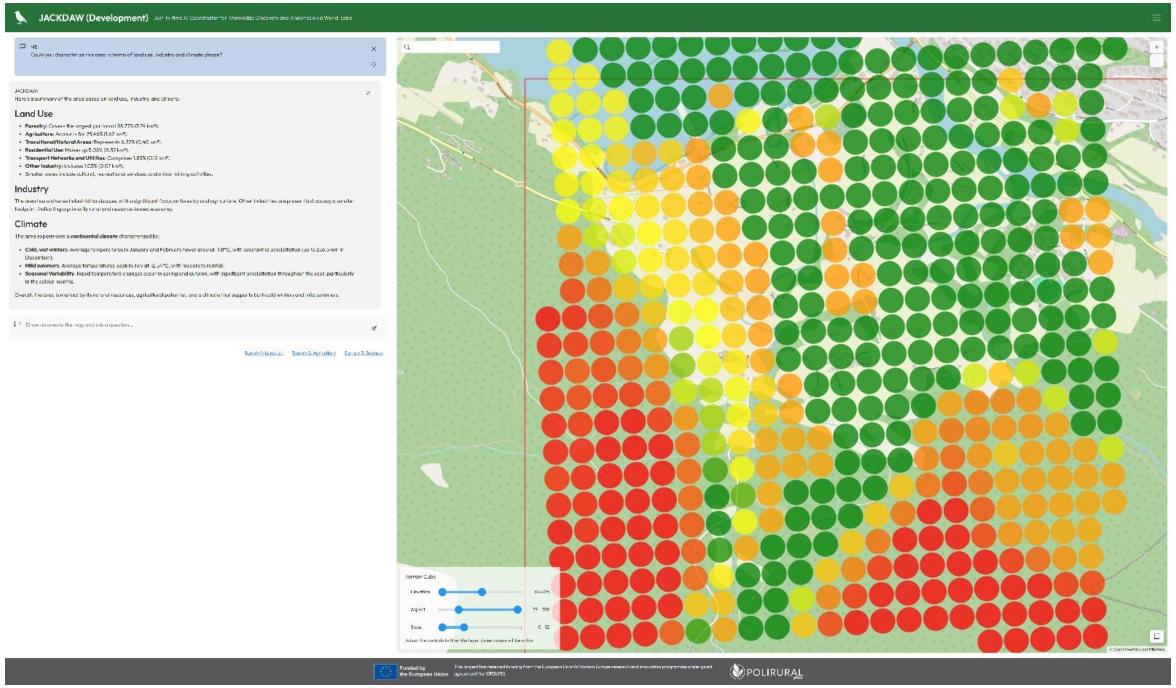
GeoRAG

- Define supported use cases and data sources: LPIS parcels, field blocks, EO-derived products, soil and terrain data, weather and climate datasets, regulatory texts and farm advisories.
- Implement a query parser that converts farmers’ and advisors’ questions into structured spatial and attribute queries (location, crop, time window, constraints).
- Build retrieval pipelines over spatial databases (e.g. PostGIS), vector embeddings and document stores to obtain relevant parcels, layers and texts for the queried area.
- Design a context assembly layer that merges spatial results (maps, statistics) with regulatory and advisory texts into a controlled prompt for the LLM.
- Integrate the GeoRAG backend with map-based user interfaces and farm management tools, and establish procedures for updating data, monitoring quality and validating outputs against expert knowledge.



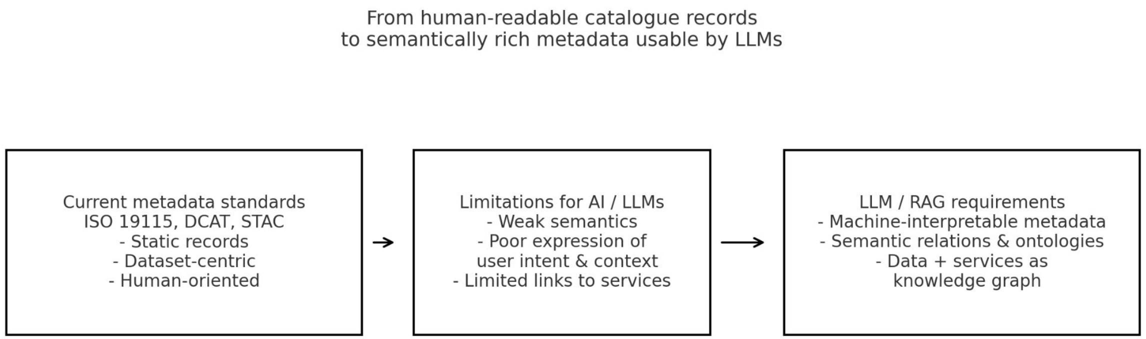
Design map-centred interfaces

- Use map-centric clients as the primary entry point for interaction with DataCubes (EO, climate, soil, yield) and parcel-based datasets.
- Allow users to formulate queries by combining spatial selection (parcel, AOI, buffer) with natural-language prompts interpreted by an LLM.
- Integrate DataCube operations (subsetting in space–time, band/indicator selection, aggregation) into LLM tool calls, so that the model can request and transform raster data on demand.
- Provide synchronised map, chart and text panels where:
- the LLM explains what has been computed and why,
- spatial results (indices, anomalies, risk maps) are visualised directly in the map,
- underlying queries and parameters remain inspectable and reproducible.
- Ensure provenance tracking: every advisory response is linked to the exact DataCube version, spatial extent, time window and processing chain used.



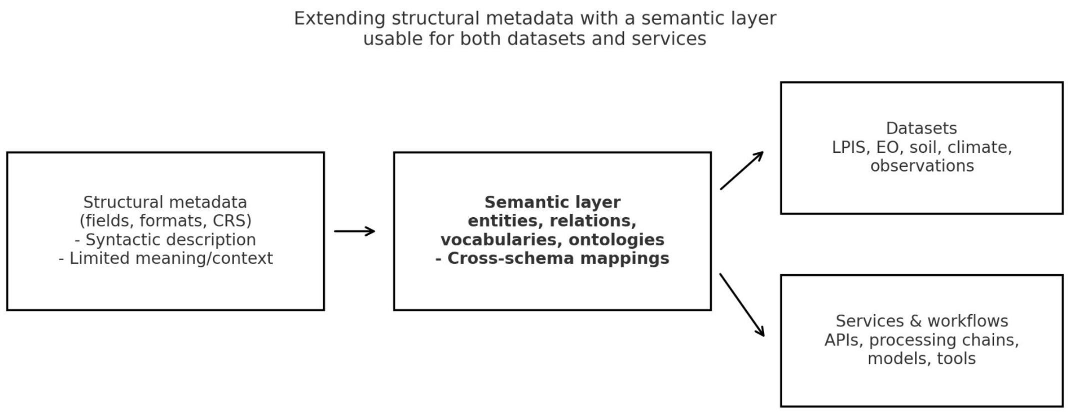
Why extend metadata for LLMs

- Existing standards (ISO 19115, DCAT, STAC) are static, dataset-centric and written for humans.
- LLMs require machine-interpretable, semantically rich descriptions of both data and services.
- Today, metadata only weakly reflects user intent, domain concepts and cross-schema relations.
- To support LLMs and RAG, metadata must become part of an operational knowledge architecture, not only a catalogue.



From structural to semantic metadata

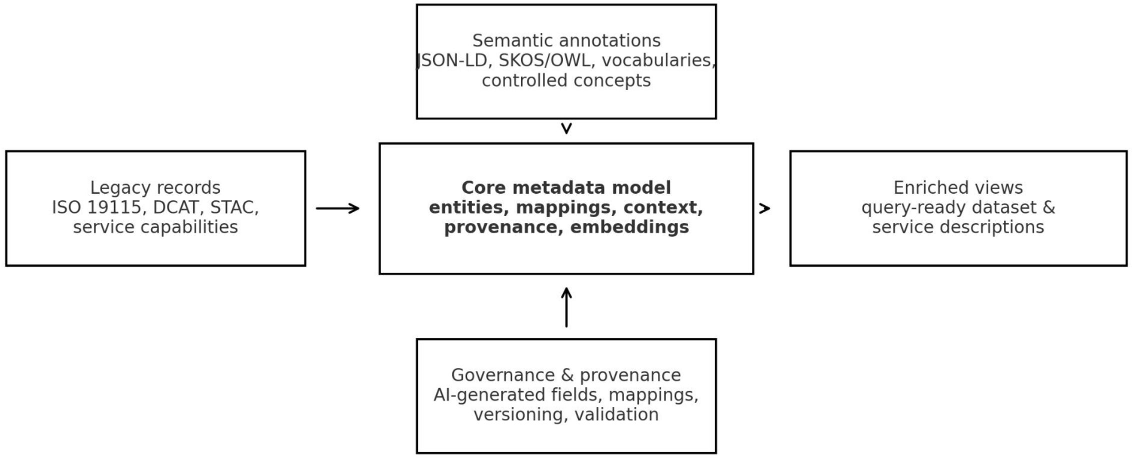
- Current records mostly expose structure (fields, formats, CRS) but not meaning or usage context.
- We need to extend metadata with semantic layers: entities, relations, vocabularies, ontologies.
- Metadata must link similar concepts across schemas (e.g. “soil moisture”, “SM”, “θv”) and clarify units, scales and validity.
- The same approach must describe both datasets and services (APIs, processing chains, models, workflows).



Semantic enrichment of data and services

- Introduce a core model for metadata entities: Dataset / Service, Schema, Mapping, Context, Provenance, Embedding.
- Attach semantic annotations (JSON-LD, SHACL, SKOS/OWL) to legacy ISO/DCAT/STAC records without breaking compatibility.
- Represent service capabilities (inputs, outputs, preconditions, costs, policies) in a structured, ontology-linked form.
- Maintain provenance and governance for all enriched elements, including AI-generated fields and mappings.

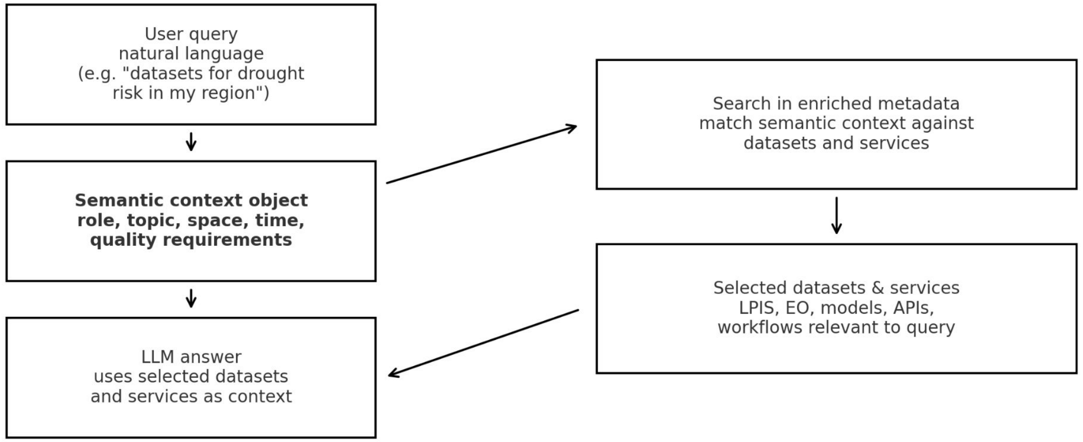
Semantic enrichment pipeline for data and services
around a core metadata model



Using enriched metadata in LLM workflows

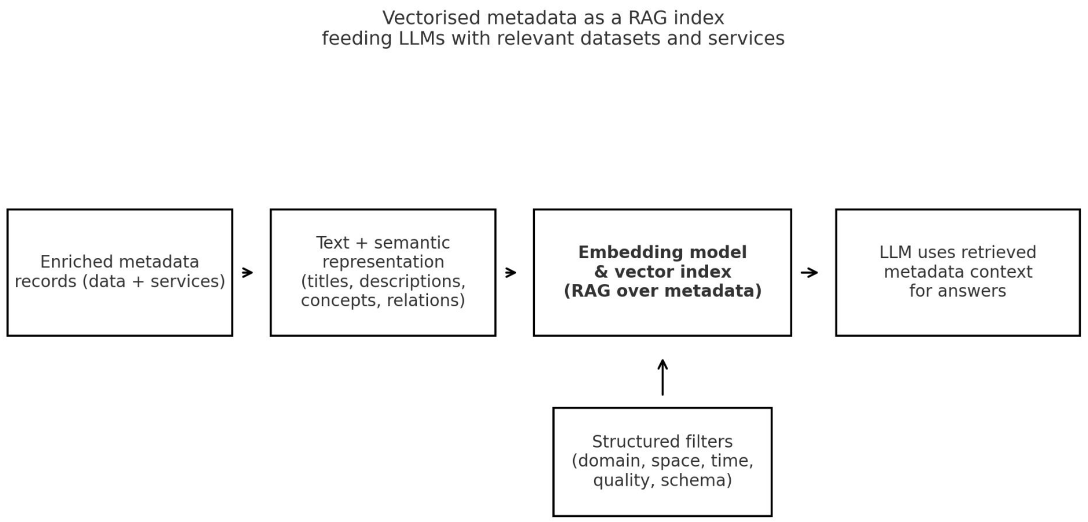
- Natural-language queries are first translated into a semantic “context object” (role, topic, space, time, quality).
- The system searches enriched metadata to find relevant datasets and services matching this context.
- LLMs exploit semantic links and mappings to bridge heterogeneous schemas and catalogues.
- The result is an adaptive metadata view tailored to the query, ready for downstream data access and processing.

Enriched metadata used as an intermediate semantic layer between queries and LLM answers



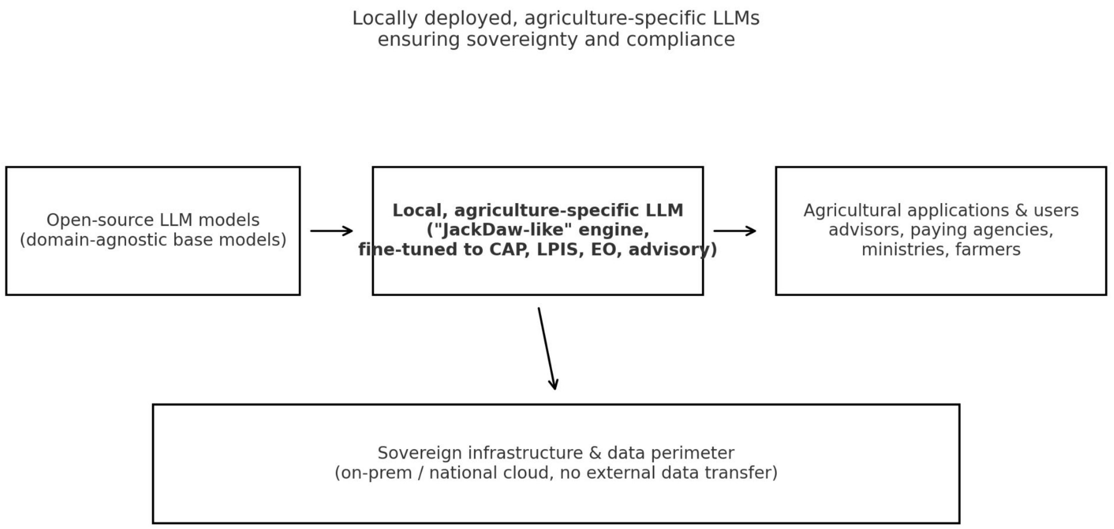
Vectorised metadata for RAG and LLMs

- Final step: convert enriched metadata (titles, descriptions, semantic tags, relations, usage notes) into vector embeddings.
- Build a RAG index over these embeddings, combined with structured filters (schema, domain, spatial/temporal coverage).
- LLMs then query the vectorised metadata directly, retrieving the most relevant datasets and services as context.
- This closes the loop: extended, semantic metadata becomes a first-class input to LLMs, enabling explainable and interoperable AI workflows.



Deploy and optimise locally running LLMs

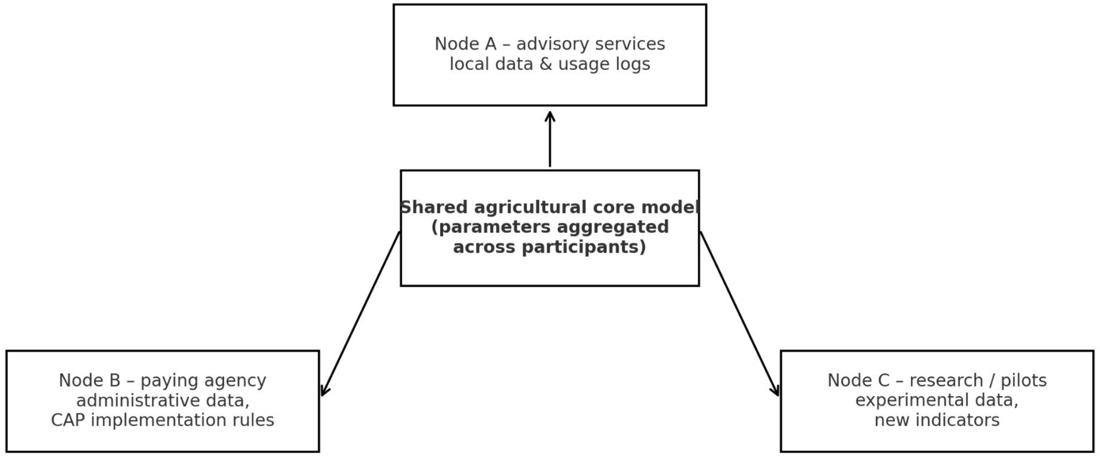
- Objective is not a single “one-size-fits-all” platform, but domain-specific solutions tailored to agriculture.
- A growing ecosystem of open-source language models can run fully on local or sovereign infrastructure, without sending data to external providers.
- These models can be adapted and fine-tuned to agricultural terminology, CAP rules, LPIS structures, EO products and advisory workflows.
- Locally deployed LLMs serve as the core of a “JackDaw-like” engine, configured for specific user groups (advisors, paying agencies, ministries) and integrated with internal data assets.
- This approach supports data sovereignty, compliance with legal and contractual constraints, and controlled evolution of the system in line with policy and organisational needs.



Federated learning – next development step

- Shift from isolated local models to collaboratively trained models across multiple organisations (advisory services, paying agencies, research institutes).
- Use federated learning to update shared model parameters without centralising raw data, preserving data sovereignty and confidentiality.
- Allow domain-specific specialisation: global “core” agricultural model plus local adaptations for national CAP implementations, languages and farming systems.
- Integrate federated training with the existing GeoRAG and metadata architecture, so that improvements in one node benefit others while respecting legal constraints.
- Establish governance, monitoring and evaluation procedures (participating nodes, update policies, validation datasets) to ensure robustness, transparency and compliance.

Federated learning for agricultural LLMs
shared model, local data sovereignty



Model is trained locally in each node; aggregated parameters
are merged in the shared core model. Raw data remain in nodes.

Next steps and opportunities for cooperation

- In the rest of November and December we will organise a series of detailed technical training sessions on the individual components of the presented architecture.
- Information and registration will be announced on the project website: https://www.poliruralplus.eu/
- The trainings are primarily targeted at winners of our open calls, but participation will be open to the wider community.
- All sessions will be recorded and subsequently made available via YouTube.
- Interested organisations and experts are invited to join these sessions as a starting point for more structured cooperation.

charvat@plan4all.eu

Name of the project:

Fostering Sustainable, Balanced, Equitable, Place-based and Inclusive Development of Rural-Urban Communities' Using Specific Spatial Enhanced Attractiveness Mapping ToolBox

Thank You for Your Attention

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101136910