OGC® DOCUMENT: 24-02#

External identifier of this OGC® document: http://www.opengis.net/doc/PER/{shortname of project}

Open Geospatial Consortium

MOVING BEYOND SDI TO GKI TO BRING GEOSPATIAL AWARENESS TO LLMS TO SUPPORT CLIMATE AND DISASTER RESILIENCE

ENGINEERING REPORT

FINAL

Submission Date: 2024-07-17
Approval Date: 2025-01-02
Publication Date: 2025-01-02
Editor: Nathan McEachen, Justin Lewis

Note

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

CONTENTS

LIST OF FIGURES

1

OVERVIEW

1 OVERVIEW

One of the key objectives of the Open Geospatial Consortium (OGC) Climate and Disaster Resilience Pilot 2024 (CDRP) is to encourage innovations that can achieve scale for emergency management analysis, decision support, and preparedness. Multiple and interrelated challenges across sectors impacted by climate change can collectively produce a larger negative effect. However, there are a number of barriers to achieving cross-sectoral collaboration and analysis at needed scale. In addition to the need for more scientific knowledge, the information systems, data, and data management practices used to address these challenges are siloed, making it difficult to share knowledge and understand the Cumulative Effects (CE) and regional impacts due to varying geography, environments, and populations.

Generative AI advances, such as large language models (LLM), show promise for achieving scale to create financially sustainable and inclusive solutions, including the often overlooked “last mile” at the base of the pyramid to serve the needs of the poorest and smallest population groups. Although LLMs can possess a high-level of general knowledge, they often lack access to the cross-sectoral, domain-specific, and new knowledge needed to answer ad-hoc questions to “what-if” scenarios useful for understanding the CE of interrelated challenges.

We highlight the shortcomings of supplier-driven and human-readable spatial data infrastructures (SDI) for facilitating cross-sectoral collaboration and outline findings from previous and current work for how a geospatial knowledge infrastructure (GKI) can achieve needed scale by automating knowledge integration for collaboration and leveraging generative AI advances. A transition to GKIs would support CDRP objectives of scalability by providing a means for representing geoscience outputs as spatial knowledge representations (i.e., geosemantics) that can be integrated automatically by location to support on-demand analysis of climate disaster CE. We describe how this can be accomplished by making spatial knowledge graphs (SKG) between organizations interoperable and how to manage the propagation of change between interlinked graphs to provide LLMs with cross-sectoral geospatial awareness as it changes over time.

2

THE MOVE BEYOND SDIS TO GKIS AND LEARNINGS FROM RELATED CLIMATE RESILIENCY WORK

2 THE MOVE BEYOND SDIS TO GKIS AND LEARNINGS FROM RELATED CLIMATE RESILIENCY WORK

The United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) has identified that SDIs are fundamentally inadequate for supporting the cross-sectoral collaboration needed to address the multiple and interrelated challenges humanity faces, as they are supplier-driven data networks and services that are single directional and created for a general market. As such, they are not sophisticated enough to allow for processing random, unpredictable, and context-dependent queries across multiple domains.

The momentum behind the 'beyond SDI' movement is the realization that 'an ever-ready plethora of geospatial data to answer individuals' questions in real time,' is not achievable with the current human-readable SDI paradigm. Rather, a geospatial knowledge sharing infrastructure is needed that is client-driven by on-demand questions of individuals and the relationships between those questions across domains. In a GKI, answers to questions can be published as knowledge services using machine-readable interfaces so that they can be integrated with other published knowledge to answer different questions in an ad-hoc and client-driven manner (UN-GGIM, 2022).

Developing a GKI for the U.S. Army Corps of Engineers (USACE) to support climate resiliency and project planning revealed the following insights. Firstly, it is important to understand the transitive relationships between networks of geographic objects (geo-objects) according to their related geographic context for CE analysis. For example, a flood model can predict areas that will be impacted by a flood event but it does not necessarily understand the CE (i.e., downstream impact) across different sectors, programs, environments, or physical infrastructures that are not captured in the model. Secondly, understanding the downstream impact requires integrating geographic information from multiple organizations. This is already a challenge within organizations such as USACE and other U.S. agencies. For example, how will healthcare delivery be affected if a health facility is damaged? What part of the population will be especially vulnerable? Thirdly, the SKG is another silo of information that will need to be maintained. This was identified as one of the biggest challenges to sustainability, as dependencies between independently developed graphs change over time and they can be prohibitively time consuming to maintain.

3

BRINGING KNOWLEDGE TO LARGE LANGUAGE MODELS

3 BRINGING KNOWLEDGE TO LARGE LANGUAGE MODELS

Although LLMs can possess a high-level of general knowledge, they lack access to cross-sectoral domain-specific and new knowledge. Ideally, LLMs should be able to provide insight into the complex and nuanced needs of individual locations for not just climate and disaster resilience, but to also support broader public health efforts. Despite the rapid increase in LLM performance, progress may plateau in a short period of time due to running out of training data. It is possible that the estimated stock of human-generated public text is about 300 trillion tokens, which could be fully utilized sometime between 2026 and 2032, or possibly earlier. This limitation may be overcome by incorporating more than just unstructured text as training data (Villalobos et. al 2024).

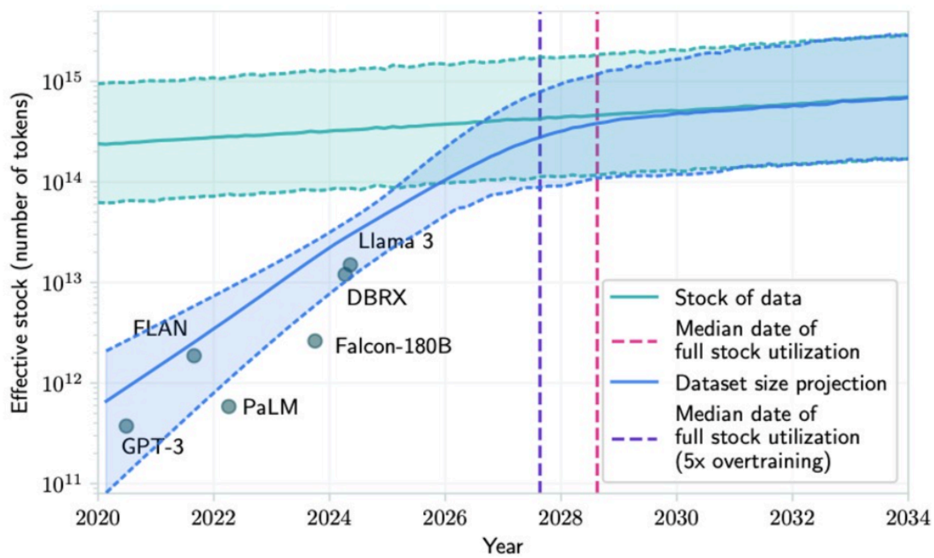

Figure 1 — LLMs may run out of unstructured training data soon https://doi.org/10.48550/arXiv.2211.04325

There is useful information from SDIs and geoscience outputs that, if made available as machine-readable knowledge, could bring valuable geospatial awareness to LLMs. According to Pan et. al, a Knowledge Graph (KG) is a data structure that could be used for this purpose. Whereas LLMs have implicit knowledge, KGs can model explicit and domain-specific knowledge to provide traceability and reduce indecisiveness and hallucinations. Methods for utilizing LLMs and KGs together is a developing field of research, but some approaches that are standing out include: 1) using KGs to train LLMs, 2) using retrieval augmented generation (RAG) to reference KGs to ensure an LLM has access to external and verifiable facts (to reduce hallucinations), and 3) using LLMs to convert natural language questions into a graph query language (Pan et. al 2023).

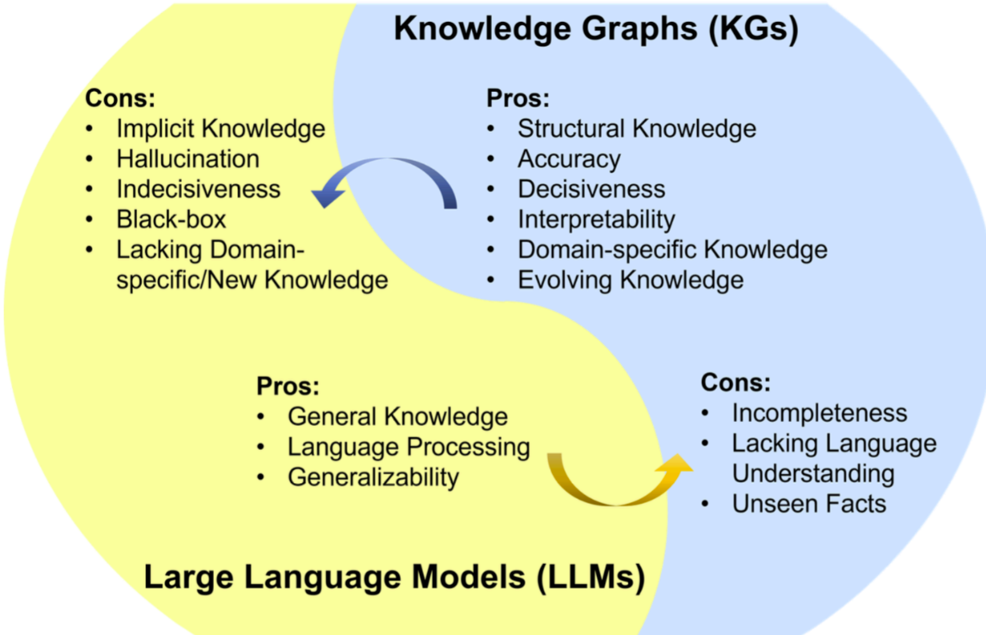

Figure 2 — Synergy between KGs and LLMs https://ieeexplore.ieee.org/document/10387715

4

SPATIAL KNOWLEDGE GRAPHS AND FEATURE-LEVEL METADATA

4 SPATIAL KNOWLEDGE GRAPHS AND FEATURE-LEVEL METADATA

How can data from SDIs and geoscience outputs across domains be represented as SKGs? In order for LLMs to have geospatial awareness down to small levels of spatial granularity, such outputs across a knowledge sharing ecosystem need to be findable, accessible, interoperable, and reusable (FAIR).

SDIs are implemented using mature spatial data management tools, systems, and APIs with metadata describing datasets rather than features. Geospatial knowledge management tools, standards, and practices with feature-level metadata are just emerging. Whereas data constitutes facts and figures without context for a specific question, information is created when data are processed, organized, or structured to add value or meaning. Knowledge is putting information into context to answer a specific question. One of the objectives of the CDRP is to identify needs and gaps in the value chain from data to Analysis Ready Data (ARD) to Decision Ready Indicators (DRI) for decision support. How do we extend this to capture knowledge in a machine interoperable way using a GKI so that an LLM can provide insight across domains at the feature level?

Tabular geospatial data structures, such as database tables and map layers, represent a set of features of a given feature class. However, they are not optimized to manage or represent how individual features across multiple feature classes relate to each other. We will use a simplified example to determine the CE of breached levees along waterways inspired from our findings working with USACE to build a GKI. For instance, a layer of *levees* does not capture which levees are within a given distance of structures. To answer that question, one would need to obtain a dataset of structures and perform some geometry arithmetic. However, the question that really needs to be answered is which structures are at risk of flooding due to a levee breach, but that does not always correspond to a fixed distance from the levee due to topography. One would then need to obtain the *leveed area* layer that was created by domain experts as a DRI containing geometries representing areas that will be flooded due to a breach. By using geometry arithmetic the link can be established between levees and structures that are at risk of flooding.

In order for this information to be findable using a SDI, metadata for a data source that contains attributes indicating which structures are at a flood risk due to a levee breach, and possibly a reference to the corresponding levee, need to be published. But there are a number of shortcomings with regard to the principles of FAIR. Unless there is a way to establish a single source of truth for feature identity across the ecosystem, then the referenced structures will not automatically be interoperable with other datasets that reference the same structures. Consequently, asking questions involving flooding and other domains would require a separate data integration effort. Additionally, the metadata should enable the ability to determine CE or downstream effects that are needed for answering “what if” scenarios. For example, which populated areas outside of the flooding event will have disrupted access to healthcare? That would require obtaining spatial tables for populated areas, hospitals, and health catchment areas, which is the area a hospital serves. Each question and layer of indirection linking datasets for CE analysis require separate integration efforts. An SDI also does not provide a convenient mechanism for publishing answers that may be reusable, nor does it make clear the transitive references to location across multiple published datasets. The diagram below illustrates the relationships between the tables and columns in the breached levee example. An SDI does not have sufficient metadata to support automatic discovery and traversal of related tables and columns to answer a question within a given domain. Rather, those relationships need to be discovered manually.

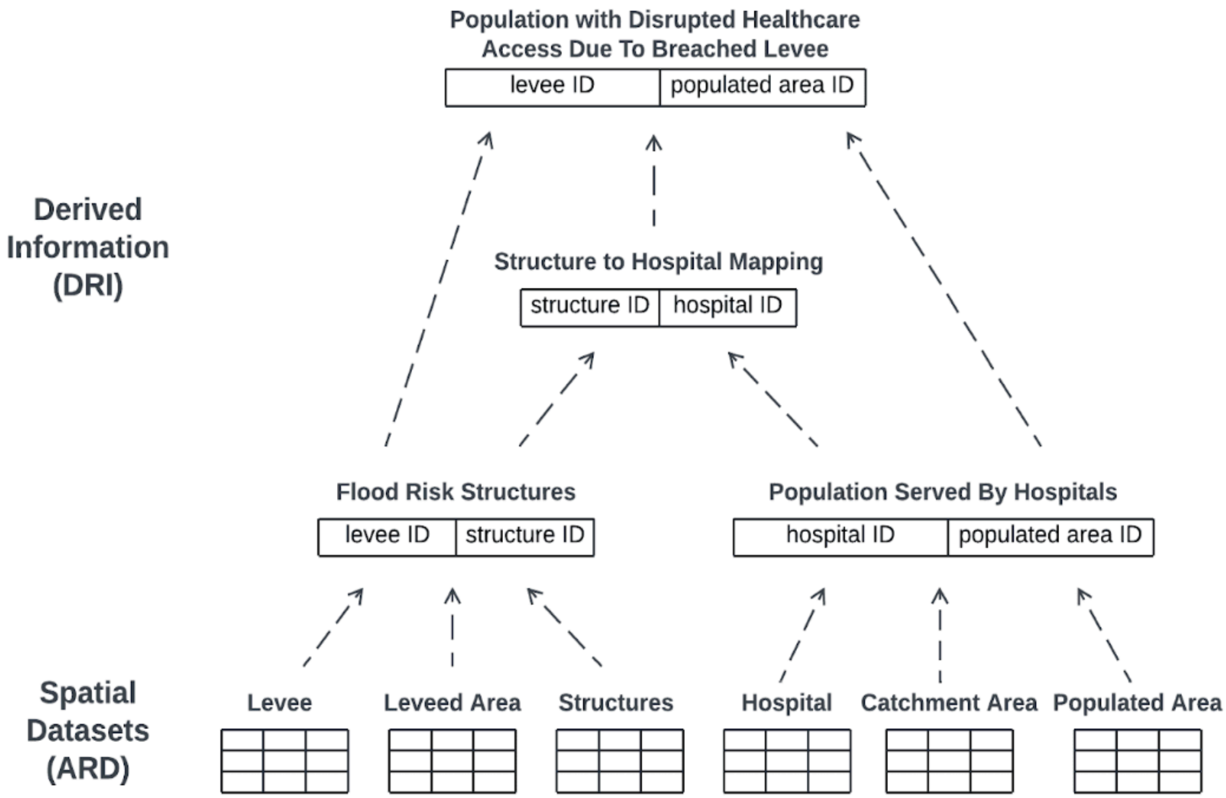

Figure 3 — SDIs do not automate the discovery of transitive relationships between features in a given domain to answer “what-if” questions

5

IMPLEMENTING A GKI USING INTERLINKED SKGS

5 IMPLEMENTING A GKI USING INTERLINKED SKGS

A key aspect of a GKI architecture is that answers to previously asked questions can be published in a machine-readable format so that they can be incorporated into answering other questions. Interoperability by location and for the correct period of time can be enabled across an ecosystem through the use of a common geo-registry to facilitate the single source of truth for geo-object identity, classification, geometry, and relationships to other geo-objects as they change over time (Health GeoLab 2022). In our previous work, we outlined how to build a GKI using interlinked SKGs maintained by the spatial knowledge mesh architecture pattern. This included a metamodel with properties that were identified as necessary for publishing spatial knowledge graphs as products interoperable by common geography (McEachen and Lewis 2023). The figure below is a simplified UML diagram of the metamodel for creating *geo-ontologies* that define classes in a SKG.

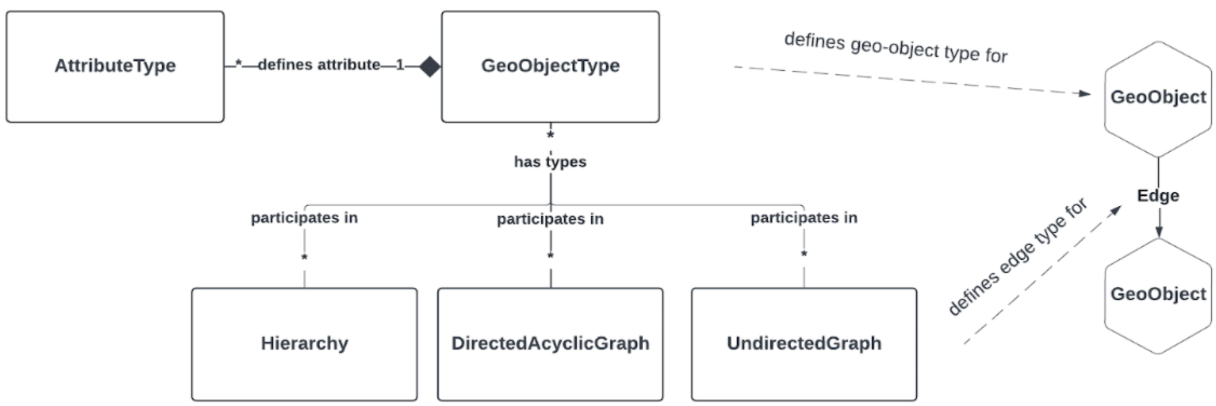

Figure 4 — Simplified metamodel for creating geo-ontologies that define classes for interoperable spatial knowledge graphs

In a spatial knowledge graph, geo-objects are nodes that represent features, whose classes are defined by geo-object types in a geo-ontology. Geo-objects are connected by edges whose types represent different semantic spatial relationships. Data commonly need to be copied from external sources in order to support analytics efforts. However, these copies should not become silos. Rather, they should function as database materialized views that maintain an association with the source data. To that end, we identified the following machine-readable metadata properties for published SKGs:

Authoritative source: Even as data are copied and linked with other sources, the authoritativeness should always be explicit and unambiguous.

Period of validity: The period for which data values are valid needs to be explicit and unambiguous. Otherwise, a temporal mismatch can occur when merging SKGs multiple sources but for incongruent periods of time, which will likely result in incorrect analysis.

Version: Metadata should explicitly capture the version of published SKGs. Periods of validity can have multiple versions.

Formally defining feature classes as geo-object types and their semantic relationships using a geo-ontology provides metadata that can be referenced by a knowledge sharing ecosystem for interoperability. The diagram below depicts a geo-ontology that defines a SKG to capture the semantic geospatial relationships of interest for determining flood CE related to our work with USACE. By using the spatial knowledge mesh architecture, which is an extension of a data mesh, organizations can independently develop their own geo-ontologies that reference geo-object types curated by the authoritative source. This metadata will allow graphs to automatically receive updates as referenced dependencies change over time (McEachen and Lewis 2023).

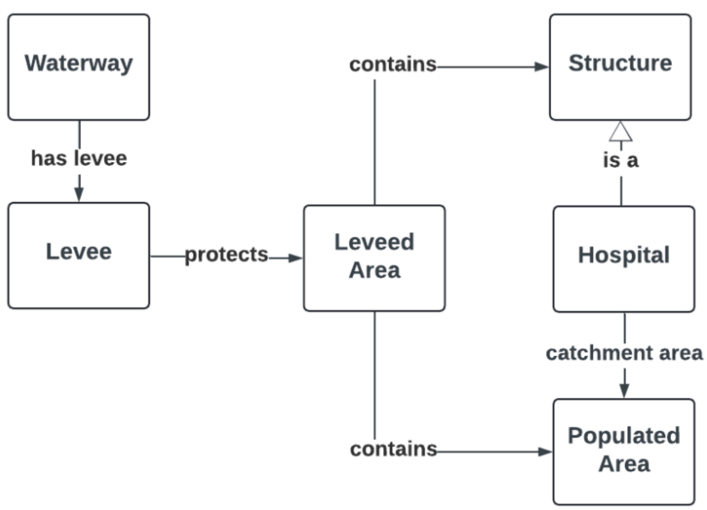

Figure 5 — A simplified geo-ontology defining the SKG in the USACE CE analysis example

The diagram below illustrates how one can associate tabular datasets with their corresponding geo-object type and edge definitions. The *Leveed Area* table has a reference to levees that map to the edge type *protects*. The calculation to determine structures at risk of flooding is mapped to the *contains* relationship connecting *Leveed Areas* and *Structures*.

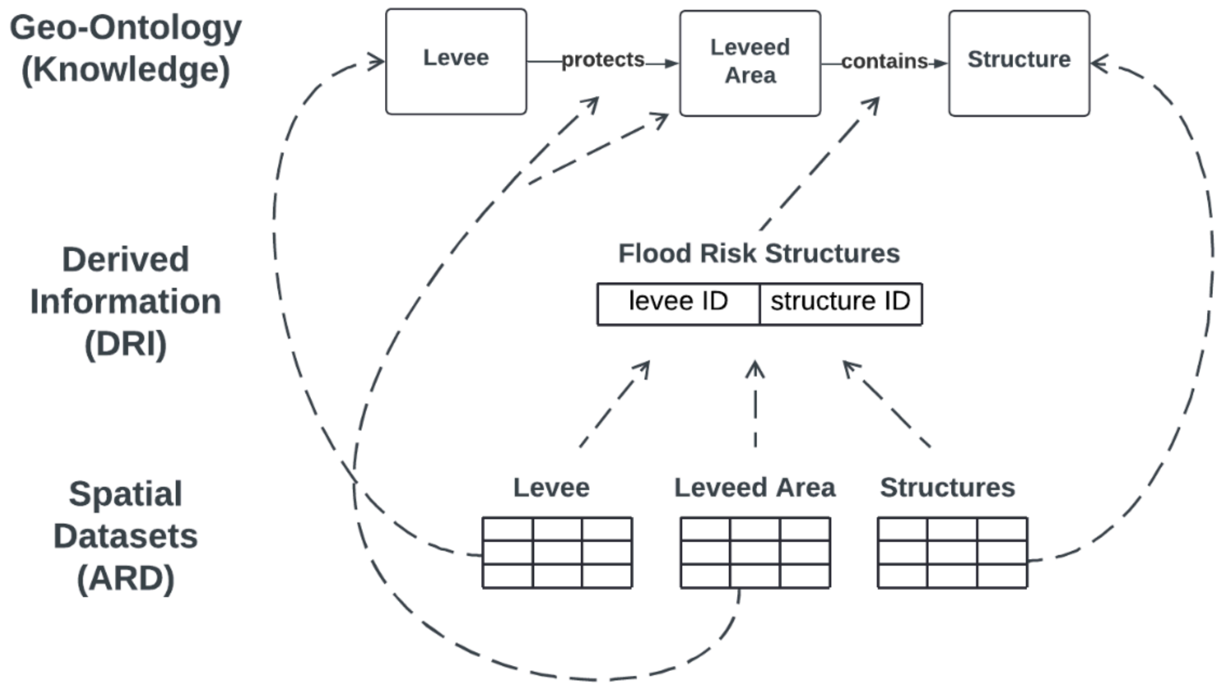

Figure 6 — Defining a GKI by mapping tables and columns to a geo-ontology so that relationships between features are discoverable

Once data and derived information are expressed in a spatial knowledge graph, whether it is mapping to the geo-ontology or storing the information in a spatial knowledge graph, the interlinked relationships can be explored to discover CE. By referencing common geo-ongologies and leveraging a common geo-registry to ensure that common geographies are semantically equivalent, then analytics outputs captured in SKGs can be interoperable using machine-to-machine readable interface to facilitate client-driven queries as envisioned by the UN-GGIM GKI concept (McEachen and Lewis 2023).

The following diagram depicts a SKG defined by the example USACE geo-ontology. By populating the graph with ARD and DRI, the transitive relationships between geo-objects (i.e., geo–features) are discoverable. In this case, although *Populated Area 2* is not impacted by the breached levee but has nonetheless lost access to healthcare services, as it is in the catchment area of a hospital that has been flooded. Also depicted is how the concept could be extended to determine disruptions due to compromised transportation infrastructure.

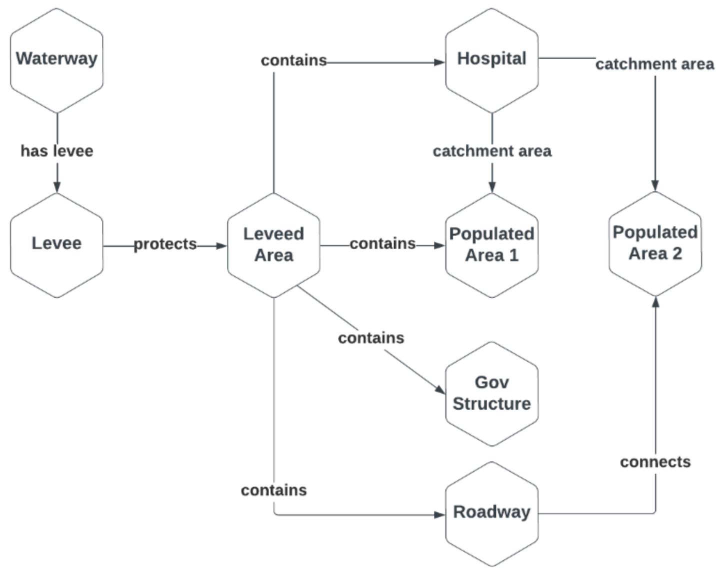

Figure 7 — SKG for exploring the CE of a breached levee

6

AUTOMATING PROPAGATION OF CHANGE BETWEEN INTERLINKED SKGS

6 AUTOMATING PROPAGATION OF CHANGE BETWEEN INTERLINKED SKGS

Modeling everything of interest relative to a feature or a network of features for analysis can be achieved by merging SKGs curated from different organizations that normalize semantic properties on common geographies. Our work with USACE and the work of the CDRP participants helped us to identify the need for additional metadata at the geo-object (i.e., geo-feature) level to help automate the merging of independently developed SKGs across organizations and domains. Once semantic identity is established, then the same geo-object referenced from multiple SKGs will be represented as one node in the merged graph. Without this, multiple nodes can appear in the graph for the same geo-object as different features. When a geo-object exists in multiple SKGs, a precedence is needed to determine how the semantic properties will be represented in the merged graph. Our work to date indicates this should be based on the period of validity and version. Each geo-object in the graph should have a reference to the metadata of the SKG it belongs to in order to establish merging precedence. In the diagram below, SKGs A, B, and C are merged into a single graph. All graphs have the same period of validity, so there is no temporal conflict. SKG B and C both reference Census Block 1. However, SKG C references a newer version, so it takes precedence in the merged graph, SKG D, which now has semantic spatial properties from two domains.

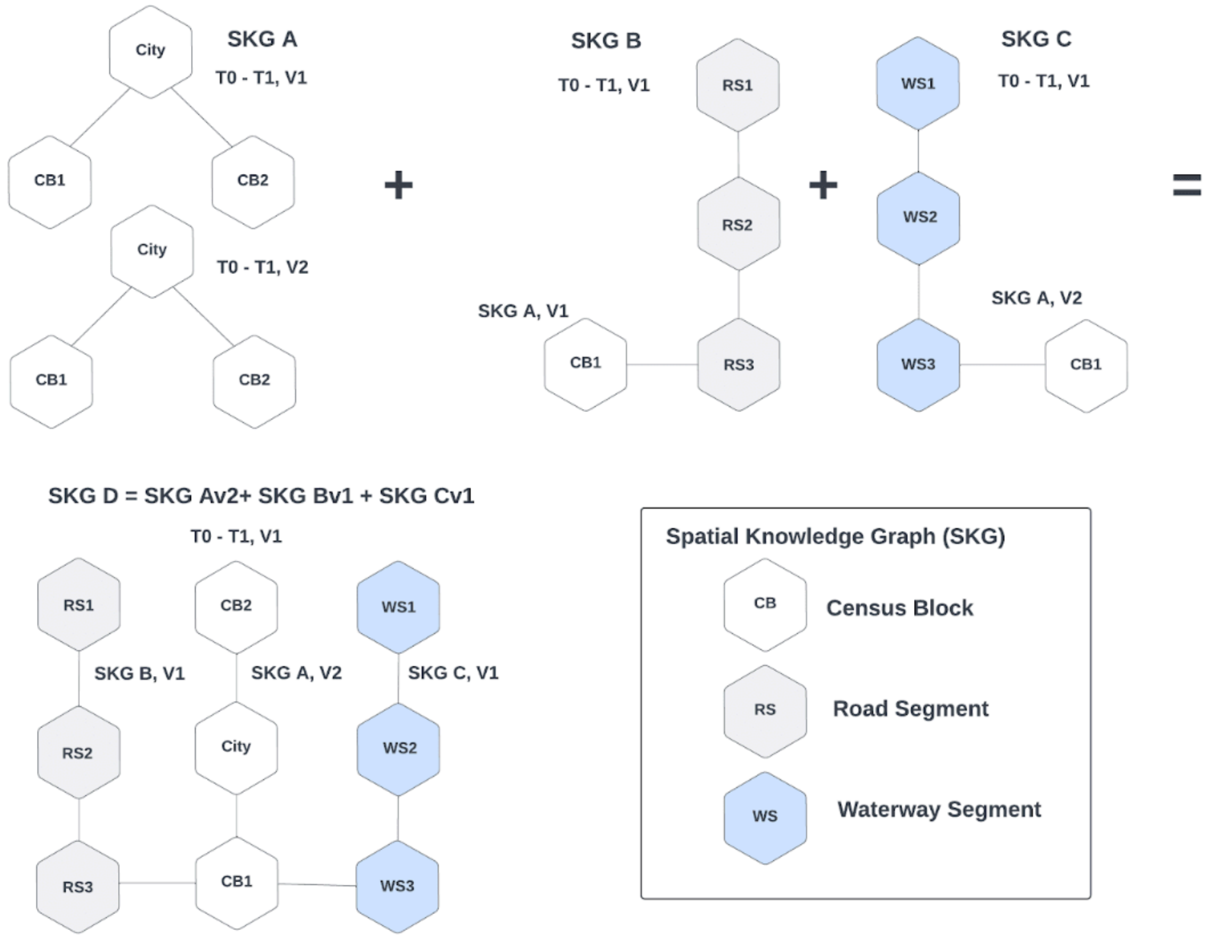

Figure 8 — Merging of two SKGs from different domains that reference common geographies

7

FUTURE WORK

7 FUTURE WORK

Possibilities of adapting this geospatial knowledge sharing approach with existing standards should be explored. OGC Rainbow (https://www.ogc.org/resources/rainbow/) shares the concept of using a metamodel for defining semantic structures, but does not support versioned periods of validity and directed and undirected semantic edge types. Additionally, it does not address managing dependencies between interlinked knowledge products within a data mesh. Synergies with using a spatial knowledge mesh as a training input and the Training Data Markup Language (https://www.ogc.org/standard/trainingdml-ai/) would also be worth exploring. The dataspaces concept is gaining traction as a way to facilitate collaboration but implements semantic interoperability on an as-needed basis. A downside of this approach is that support for ad-hoc and client driven queries across domains would be limited to the datasets that had already been integrated. Could the principles of a GKI minimize this integration overhead? There are some conceptual similarities with the OGC paper titled *Modernizing SDI: Enabling Data Interoperability for Regional Assessments and Cumulative Effects CDS* that include interoperability and geosemantics, push-based communication flows as a means for ensuring propagation of change, and more support for client-driven and on-demand queries (Thomas and Lieberman 2021). Our algorithm for merging spatial knowledge graphs will need to be tested and adapted to identify and resolve conflict scenarios including incompatible periods of validity.

8

CONCLUSION

8 CONCLUSION

A GKI implemented with interlinked SKGs enables cross-sectoral collaboration by automating the integration of knowledge representations by common geographies. The significance for scalability for the CDRP is that domain experts for flood, mudslide, or heat island analytics should not have to shoulder the responsibility of determining cumulative effects and downstream impacts. Rather, other entities can independently merge climate with infrastructure, public health, and transportation graphs to create fit-for-purpose models for addressing the nuanced needs of local geographies as geospatial awareness for LLMs.

We would like to thank USACE for supporting the effort to operationalize a GKI and to thank the OGC community of partners for their shared learnings that helped guide our approach.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] UN-GGIM, 2022. Development of Contextual Understanding, Information, and Analytics Towards Determining the National Geospatial Information Ecosystem, https://ggim.un.org/meetings/GGIM-committee/12th-Session/documents/Future_Geospatial_Information_Ecosystem_Discussion_Paper_July2022.pdf. (4 April 2024)

[2] Thomas, R., Lieberman, J., (2021) Modernizing SDI: Enabling Data Interoperability for Regional Assessments and Cumulative Effects CDS, Reference number: OGC 21-013, http://www.opengis.net/doc/PER/Mod-SDI-CDS (12 July 2024)

[3] Health GeoLab Collaborative, 2022. HIS geo-enabling: Guidance on the establishment of a common geo-registry for the simultaneous hosting, maintenance, update, and sharing of master lists core to public health (Version 2.0). https://healthgeolab.net/DOCUMENTS/Guidance_Common_Geo-registry_Ve2.pdf. (4 December 2022)

[4] McEachen, N. and Lewis, J.: ENABLING KNOWLEDGE SHARING BY MANAGING DEPENDENCIES AND INTEROPERABILITY BETWEEN INTERLINKED SPATIAL KNOWLEDGE GRAPHS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-4/W7-2023, 117–124, https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-117-2023, 2023.

[5] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, https://doi.org/10.1109/TKDE.2024.3352100

[6] Villalobos, P., Ho, A., Sevilla, J, Besiroglu, T., Heim, L., Hobbhahn, M., (2024) Will we run out of data? Limits of LLM scaling based on human-generated data. https://doi.org/10.48550/arXiv.2211.04325