

Integrating Multi-Modal Spatial Data using Knowledge Graphs – a Case Study of Microflora Danica

Mads Corfixen¹, Thomas Heede¹, Tomer Sagi¹, Mads Albertsen², Thomas D. Nielsen¹ and Katja Hose^{3,1}

¹Department of Computer Science, Aalborg University, Aalborg, Denmark

²Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

³Institute of Logic and Computation, TU Wien, Vienna, Austria

Abstract

Integrating semantically related, multi-modal, heterogeneous data sources is challenging, especially if one of the modalities includes spatial data, such as field measurements organized in geographical grids. Since geographical grids can have different rotations, be translated along one or more axes, or have different resolutions, a particular challenge when integrating such data is to reduce the information loss from projecting different grids into a common format. In this paper, we study this problem and sketch a method for integrating such spatial data using knowledge graphs. We discuss this solution in the context of a real-world use case, where we integrate geographically annotated microbial data (Microflora Danica) as well as environmental data to enable joint analysis. The first results of our experiments show that our method reduces the information loss compared to baseline methods.

Keywords

Spatial Data Integration, Knowledge Graph, S2 Geometry

1. Introduction

Integrating and linking semantically related heterogeneous and multi-modal data is an important task that is often very challenging [1, 2]. While many works have focused on integrating tabular data, defined over a fixed relational schema, recent years have seen a diversification into combining multi-modal data (such as semi-structured data from different modalities) [3] and data with a flexible schema [4].

The need for data integration appears in a multitude of domains. One such domain is the study of microbiomes, i.e., the interaction between microbes and the environment they inhabit. Understanding these interactions is essential for solving current and future environmental challenges [5]. Such data is inherently multi-modal because the DNA sequence data is of a different modality than the semantic relations between microbial species and the spatial

SeWebMeDA-2024: 7th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 26, 2024, Hersonissos, Greece

✉ maco@cs.aau.dk (M. Corfixen); thhe@cs.aau.dk (T. Heede); tsagi@cs.aau.dk (T. Sagi); ma@bio.aau.dk (M. Albertsen); tdn@cs.aau.dk (T.D. Nielsen); katja.hose@tuwien.ac.at (K. Hose)

🆔 0009-0004-7237-9694 (M. Corfixen); 0009-0004-7934-1293 (T. Heede); 0000-0002-8916-0128 (T. Sagi); 0000-0002-6151-190X (M. Albertsen); 0000-0002-4823-6341 (T.D. Nielsen); 0000-0001-7025-8099 (K. Hose)



© 2024 Copyright © 2024 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

information about their environments. Spatial information in general often refers to vector or raster data. Vector data consists of (x, y, z) -coordinates associated with real-world measures of a specific area, where the area's shape is defined by vector coordinates and can take the form of points, lines, polylines, and polygons. On the other hand, raster data is a defined grid over an area where each grid cell has a measurement value. The raster file itself is associated with information such as the geographical location and the height and width of the grid.

One method to facilitate the integration of multi-modal data is through a knowledge graph (KG) [6]. An advantage of using KGs for data integration is that they can flexibly support multiple modalities [7]. KGs can cover strings, images, and audio, as well as links between them, effectively allowing heterogeneous data to be mapped directly without requiring any transformations to conform to strict schemas required by other data integration solutions. Another advantage is that KGs are associated with ontologies that clearly define the semantics of the concepts in the data sources. In contrast, table and column headers in a relational database are often only clearly defined in the mind of the designer [8].

In this paper, we present a case study for integrating spatial heterogeneous multi-modal data sources using a KG. The case study is based on the Microflora Danica¹ (MfD) dataset, an archive of microorganisms in Denmark. We present the challenges and possible solutions of enriching this dataset with data from EcoDes-DK15 [9], a high-resolution dataset of ecological descriptors, and soil maps of Denmark [10]. We focus on integrating the data along the spatial dimension, as the datasets are connected through their geographical properties, i.e., GPS coordinates for the MfD dataset and spatial raster maps for the EcoDes-DK15 and soil maps. The remainder of this paper is structured as follows. First, in Section 2, we provide an overview of existing approaches for integrating multi-modal data with a spatial component. Then, in Section 3, we present the different data sources of our case study and their different modalities. Next, in Section 4, we present our KG design for spatial integration and show how it can be extended with microbial data. In Section 5, we evaluate our spatial integration approach based on the information loss and compare it to a baseline method. Finally, Section 6 concludes the paper with a summary and an outlook to future work.

2. Related Work

Data integration using KGs has recently gained attention, mainly due to their flexibility, but also because of the advantages when dealing with heterogeneous multi-modal data [6, 11]. While many studies have explored integrating or transforming spatial data into KGs, they focus mostly on vector data [1, 12, 13, 14]. Integration of raster data has not been thoroughly explored.

GeoTriples [12] proposes a method for transforming spatial data in vector format into RDF. The method extends the mapping languages R2RML² and RML³ to define rules for mapping structured and semi-structured, heterogeneous vector data to RDF graphs. They recognize that spatial data also exists in raster format but only provide a framework for transforming spatial vector data into RDF triples. TripleGeo [15] is a similar method, but the proposed framework also does not support raster data.

¹<https://www.bio.aau.dk/forskning/projekter/microflora-danica>

²<https://www.w3.org/TR/r2rml/>

³<https://rml.io/specs/rml/>

Tran et al. [16] propose an ETL process for integrating files in raster format to build an RDF triple store over a designated area. They aggregate raster cells to extract a single value for a territorial area, which leads to a loss of fine-grained information.

Zhu et al. [17] model observations as aggregations on discrete global grid systems, albeit without providing an actual framework. Moreover, they describe the integration of singular event geometries, such as wildfires, rather than integrating raster data.

Several hierarchical discrete global grids (HDGG) have been proposed to represent spatial data, including H3⁴, Bing Maps Tile System⁵, and S2 Geometry⁶. Common among them is the hierarchical division of the Earth into subsets. H3 utilizes hexagons, but, since hexagons cannot be perfectly subdivided into smaller hexagons, child cells are only approximately contained within their parent cells. Bing Maps Tile System projects the Earth onto a map; however, this projection distorts the scales in proportion to the distance to the poles. The S2 Geometry decomposes Earth into a hierarchy of cells. Instead of mapping points to a plane, S2 maps them to a perfect sphere. Since Earth is closer to being a sphere than a plane, this creates less distortion. At the top-most level of the S2 hierarchy, Earth is represented by six cells, perfectly covering the Earth, with each lower level subdividing each cell into four children.

The KnowWhereGraph (KWG) [18, 19] is a KG using S2 Geometry as the spatial component. The KWG focuses on how to model spatial data as RDF triples, but without providing a framework for mapping between raster files and S2 Geometry. Additionally, they recognize the advantages and limitations of the S2 Geometry as the spatial component but do not quantify the error. Thus, our work is based upon the KWG, using it as a spatial layer to enable spatial data integration.

3. Data Sources and Use Case Description

In this paper, we consider a use case where we want to integrate several real-world heterogeneous data sources of microbes and ecological descriptors of their habitats, exploiting the capabilities of KGs for flexible multi-modal integration of spatially related data. In this section, we describe the available data sources and how they shape our case study.

The Microflora Danica Dataset. The Microflora Danica⁷ (MfD) dataset comprises more than 10,000 measured samples collected from different sample sites in Denmark. The MfD dataset was created to provide a database of all microbes in Denmark, and contains several different modalities. Information on where each sample was taken is geographical vector data, and DNA reads can be considered as very long strings. Each sample in the MfD dataset is sequenced at least once, which is a process that constructs probable DNA reads from a sample. Through this process, each sample is associated with approximately 10 million DNA reads, R , each being a fragment of a full genome consisting of the four bases, Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), found in DNA, $R \in \{A, T, C, G\}^+$. In total, the MfD contains 28 TB of reads.

The structure of the MfD dataset is shown in Figure 1. Each sample site is described in the **Fieldsample Metadata**, which contains information such as GPS location and habitat type,

⁴<https://www.uber.com/en-SE/blog/h3/>

⁵<https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system>

⁶<https://s2geometry.io>

⁷<https://www.bio.aau.dk/forskning/projekter/microflora-danica>

with the habitat types linked to three external taxonomies. Each sample is sequenced multiple times, each sequencing being described in the **Sequencing Metadata**, linked to their related sample site through a *fieldsample_barcode*. The **Sequencing Metadata** describes how each sample was sequenced and handled in the laboratory. Each sequencing leads to a number of reads, collected in a single file we denote as **Read Data**, connected to their related sequencing through a *sequence_id*. In a **Read Data** file, there is, besides the reads themselves, a certainty measure for each base of the reads. Finally, each site was sampled as part of different projects, each project potentially collecting multiple samples. The projects are described in the **Project Metadata**, containing information on the parties responsible for the projects. Each sample site is linked to its related project through a *project_id*.

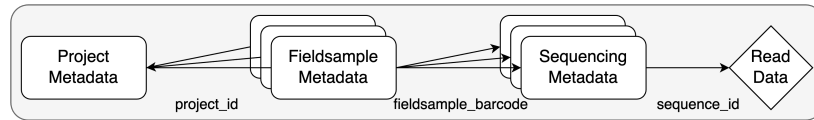


Figure 1: Overview of the Microflora Danica dataset

Besides information on the reads, sequences, and sample site, we are also provided a mapping from each sample to potential microbial species (taxons) present in that sample. Since reads are only fragments of the complete genome of a taxon, the mappings are uncertain. They are structured as shown in Table 1. Each Operational Taxonomic Unit (OTU) is a DNA sequence that encodes a specific, potentially undiscovered taxon. Each *MFD_X* column represents a *fieldsample_barcode* and shows how many reads from a sample site were mapped to that OTU.

Table 1
Mappings from sample sites to taxons

OTU	MFD_1	MFD_2	...	Kingdom	...	Species
OTU_1	0	7	...	Archaea	...	MFD_s_17257
OTU_2	633	482	...	Bacteria	...	Cornyebacterium
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Environmental Raster Files. The EcoDes-DK15 [9] and soil maps [10] datasets contain measurements of ecological descriptors across Denmark, providing us a means to quantify the microbial habitats of the MfD dataset. Both datasets are in the form of raster files, adding a new modality that needs to be integrated besides the ones present in the MfD dataset.

A raster file is a matrix organized into a grid of rows and columns of raster cells, each cell representing a geographical area. Each raster cell of a raster file is furthermore associated with a value representing information about the geographical area, such as temperature or pH. In total, the environmental data contains approximately 200 GB across approximately 100 files, each containing a different type of measure.

A raster file contains the geo-location of the upper-left cell of its matrix and a transformation matrix containing information on what translations and rotations are needed to get the rest of the spatial locations of all the other cells. In our datasets, raster cells have a resolution of 10×10 meters, and the spatial location is given on the Universal Transverse Mercator (UTM)

coordinate system. In the UTM coordinate system, each location is defined by the pair (e, n) with e , *easting*, being the distance east of the Greenwich Meridian, and n , *northing*, being the distance north of the equator.

The S2 Geometry. In order to handle the integration of raster files with potentially different transformation matrices, we integrate them into a common grid, the S2 Geometry, as all raster cells in a raster file can be mapped to a set of corresponding cells in the S2 Geometry, regardless of the transformation matrix.

The S2 Geometry is a 31-level hierarchical grid that decomposes Earth into a hierarchy of cells. At the top-most level (level 0) of the hierarchy, Earth is divided into six cells perfectly covering it, while each higher level of granularity subdivides each cell into four children, such that there are 24 cells at level 1, 96 cells at level 2, and so on.

4. Spatial Integration Approach

In this section, we present our KG design for spatial integration, demonstrating how the S2 Geometry can be adopted to independently integrate the spatial aspect of each data source. Further, we show how we extend the KG design to integrate the rest of the MfD dataset. The code is available on GitHub⁸.

Spatial Data. We build upon well-established ontologies to model the spatial dimension, as shown in Table 2. The *geo* ontology is used to model the spatial relations between raster cells, S2 cells, and MfD sample sites; the upper level ontology *oboe* is used to model the environmental observations and measurements; and the *kwg-ont* ontology is the ontology of the KnowWhereGraph.

Table 2
Ontologies for the spatial part of the KG design

Prefix	IRI
<i>geo</i>	< http://www.opengis.net/ont/geosparql# >
<i>oboe</i>	< http://ecoinformatics.org/oboe/oboe.1.2/oboe-core.owl# >
<i>kwg-ont</i>	< http://stko-kwg.geog.ucsb.edu/lod/ontology# >
<i>rdf</i>	< http://www.w3.org/1999/02/22-rdf-syntax-ns# >

We can consider a raster file as a collection of raster cells, where each raster cell is associated with at least one measurement. For raster files where their raster cells perfectly overlap one another, such as they do in our use case, each raster cell is associated with multiple measurements (pH value, soil salinity, etc.). This is modeled using the *oboe* concepts *ObservationCollection*, *Observation*, *Measurement*, and *MeasurementType*; and properties *hasMember*, *hasMeasurement*, *hasValue*, and *containsMeasurementsOfType* (see Figure 2).

To model the spatial relation between S2 Geometry and the MfD sample sites, we link each site to the S2 cell that covers it, using the GPS location obtained from the MfD dataset, described in Section 3. For the spatial relation between the raster cells and S2 Geometry, we associate

⁸<https://github.com/MicrobialDarkMatter/MfD-spatial-integration>

each raster cell with the set of S2 cells that covers it. To model these spatial relations between raster cells, S2 cells, and MfD sample sites, we utilize the *geo* ontology properties *coveredBy* and *covers*, see Figure 2, where the data from the raster files is colored in orange, the data from the MfD dataset is colored in blue, and the spatial integration layer, S2 Geometry, in bold.

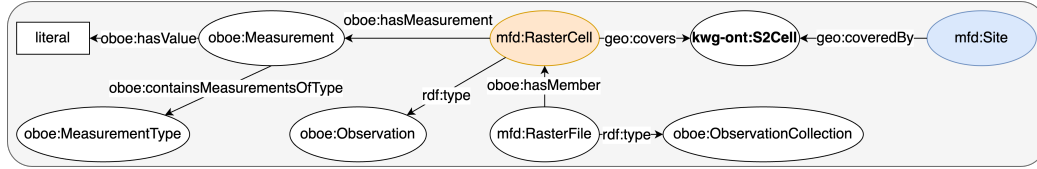


Figure 2: Central part of KG design regarding spatial data

Our current datasets do not have significant overlaps, so we do not have significant interoperability conflicts among them. However, in the future, we would like to integrate data from different sources; therefore, a semantic alignment pipeline will be needed.

Spatial Integration. The spatial aspect of each data source is mapped to the S2 Geometry. The integration can be split into two types: the vector-based GPS integration of the MfD sample sites and the area-based integration of raster files. We focus on the integration of raster files in this section, as the integration of the vector-based MfD sample sites simply requires us to integrate the given coordinates directly to the covering S2 cell.

As described in Section 3, the S2 Geometry is a hierarchy of levels determining the granularity of S2 cells. Increasing the granularity of the S2 level results in the area of each S2 cells being lowered by a factor of 4. Therefore, the S2 level affects the number of S2 cells required to cover each raster cell; a high granularity requires many S2 cells, whereas a low granularity requires fewer. Consequently, lower granularities of the S2 level increase the over-coverage, i.e., the area of S2 cells that covers an area outside their associated raster cells. The over-coverage is illustrated for two different granularities of the S2 level in Figure 3, where the S2 cells (red) are covering area outside the raster cell (green).

Ideally, a set of S2 cells covers each raster cell perfectly; however, this is not the case, and as such, choosing the highest granularity will yield the lowest possible over-coverage. The issue with this is that each level quadruples storage requirements, as each S2 cell has four child cells.

The EcoDes-DK15 and soil raster files have a resolution of 100 m²; hence, to minimize over-coverage, an S2 level of 24 is chosen as S2 cells have an area of approximately 0.3 m² at that S2 level.

Irrespective of the S2 level, S2 cells over-covering a raster cell will overlap neighboring raster cells. This is not desirable since each S2 cell should only correspond to one value per feature, e.g., we cannot associate two pH values with the same location. To address this issue, we use a majority rule; however, one could also aggregate the values, as long as they are continuous. The majority rule disassociates S2 cells from any raster cells in which their centroids are not located. However, disassociating S2 cells from a raster will introduce under-coverage, i.e., the sub-area of a raster cell that is not covered by any S2 cells.

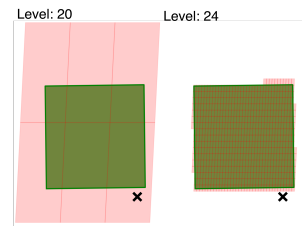


Figure 3: S2 cells covering raster cells at different levels

A potential error introduced by over- and under-coverage is integrating an MfD site to an S2 cell corresponding to an incorrect raster cell. For example, in the left of Figure 3, the sample site (cross) is linked to an S2 cell that covers the shown raster cell instead of an S2 cell covering the raster cell below it. S2 cells of higher granularity diminishes this problem.

In general, the integration of raster files is challenging due to different raster files possibly having different transformations, i.e., resolutions, rotations, and translations. However, utilizing the S2 Geometry, the integration of raster files is independent across different transformations. Thus, our approach affords the integration of arbitrary raster files.

Microbial Data. The MfD dataset contains sample-specific information such as metadata, DNA reads, sequencing metadata, sample-to-OTU mappings, and habitat information of the sample site. As the focus of this paper is on spatial integration, we omit details on this part; however, we outline the design for this integration in Figure 4. Note that each entity has properties, but due to limitations, we do not discuss metadata associated with the microbial data. Each sample site, marked in blue, has a habitat type, such as 'forests' or 'grasslands', which we map to a corresponding concept in the Environment Ontology, *ENVO* [20], marked in green. Furthermore, each sample has a mapping to an OTU, which encodes for a specific microbial taxon. These taxons are mapped to corresponding taxons in the the Taxonomy Database ontology, *NCBITaxon*⁹, marked in green. Both the habitat and taxon mappings to external ontologies require some entity resolution methods to account for, e.g., spelling errors. Finally, since the DNA reads, marked in red, take up 28 TB of storage, we do not keep them directly in the graph, but instead store a reference to an external key-value store.

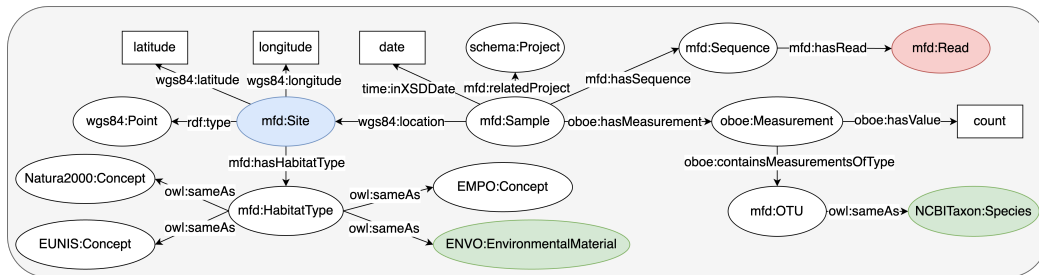


Figure 4: Extending the KG design with MfD microbial data

5. Evaluation

In this section, we evaluate the proposed framework in terms of the information loss of using the S2 Geometry as the integration layer and compare it to a baseline method. Furthermore, we highlight potential issues when using the S2 Geometry.

Information Loss. We observe two contributing factors to the information loss from our integration to the S2 Geometry. The first type of information loss is over-coverage (Equation 1), where a part of the area of the

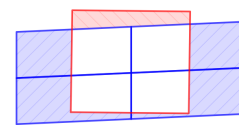


Figure 5: Over- and under-coverage

⁹<https://obofoundry.org/ontology/ncbitaxon.html>

minimal set of covering S2 cells associated with a raster cell is outside the bounds of the raster cell. The over-coverage is calculated as the proportionate area of the set of S2 cells not contained within the raster cell. The second type of information loss results from the implementation of the majority rule, and is denoted under-coverage (Equation 2), where a raster cell is only partly covered by its associated minimal set of covering S2 cells. The under-coverage is calculated as the proportionate area of the raster cell not contained within the set of S2 cells. These two types of information loss are visualized in Figure 5, with the blue area being over-coverage and the red area being under-coverage.

$$OC = \frac{|S_2 \setminus \text{Raster}|}{|S_2|} \quad (1)$$

$$UC = \frac{|\text{Raster} \setminus S_2|}{|\text{Raster}|} \quad (2)$$

Since the two contributing factors are equally unwanted, we define the information loss as the harmonic mean of the over- and under-coverage (Equation 3). Due to the nature of the harmonic mean, higher values are desirable; however, this is not the case for the over- and under-coverage calculations. Therefore, we use the complement of each coverage. This yields an information accuracy, which we subtract from 1 to get the loss.

$$\text{Information Loss} = 1 - 2 \frac{(1 - OC) \cdot (1 - UC)}{(1 - OC) + (1 - UC)} \quad (3)$$

The mean information loss for integrating with and without the majority rule is reported in Figure 6 for different S2 levels, based on random sample of 100 raster cells. Low granularity S2 levels have a mean information loss approaching 1, whereas higher granularities of the S2 level reduce the information loss. The advantage of using the majority rule is visible for higher granularities, starting from S2 level 20. However, for low granularity S2 cells, the majority rule introduces high under-coverage, as many raster cells become associated with no S2 cells, due to the centroid of large S2 cells that covers many raster cells being located only within a single raster cell.

The S2 levels 1 through 15 and 27 through 30 are not shown, as these approach an information loss of 1 and 0, respectively. At S2 level 24, where the information loss is 0.022, the information loss begins to stagnate with higher granularity of the S2 levels, which is an indication that this is a suitable level to integrate the use case raster files into.

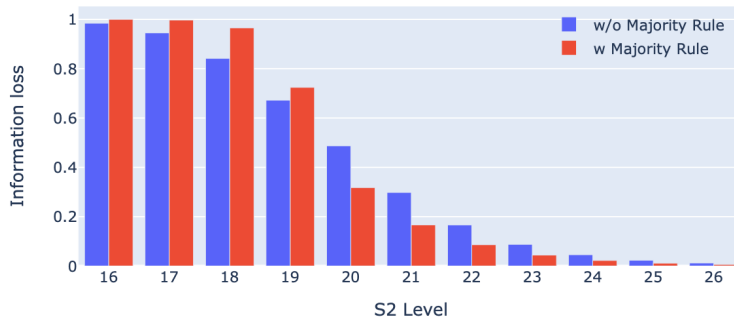


Figure 6: S2 error comparison at different hierarchy levels, with and without majority rule

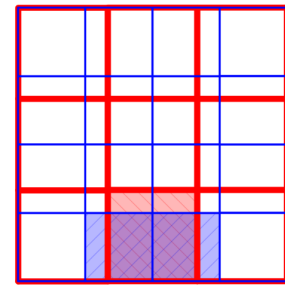


Figure 7: Upsampling and downsampling rasters of different resolutions

Using the S2 Geometry as the spatial layer is not without issues, especially when using a granularity for the S2 cells. The grid of the raster files used for this case study is approximately $35,000 \times 45,000$, resulting in 1.575 billion raster cells. For an S2 level of 24, each raster cell is linked to 625 S2 cells on average. Combining these two numbers yields a total of approximately 1 trillion triples. However, as the raster files for this case study lack measurements for raster cells located in the ocean, the actual amount of integrated raster cells diminishes by around 75%, yielding approximately 250 billion triples solely for the integration of raster cells.

Up- and Downsampling. To evaluate the gain of using the S2 Geometry as an integration layer, we compare it to using up- and downsampling for combining raster files of different resolutions. Upsampling refers to taking a set of cells at a lower granularity and then aggregating them into higher granularity and vice versa for downsampling. We exemplify the information loss from up- and downsampling via two hypothetical resolutions 10×10 and 7.5×7.5 meters, in red and blue, respectively, as illustrated in Figure 7.

In order to upsample and downsample, we define how the majority rule works in this setting. For the highlighted red cell, we see that it is covered by four different blue cells. Since only two of the blue cells have their centroids within the red cell, we downsample only into those two cells. Conversely, we upsample the highlighted two blue cells into the single red cell in which their centroids are located. We note that upsampling is more difficult than simply aggregating if we deal with categorical attributes.

An information loss of 0.5 is obtained for upsampling into the 10×10 grid and 0.298 for downsampling into the 7.5×7.5 grid. In comparison, integrating the two grids into the S2 Geometry at level 24 results in an information loss of 0.028 and 0.019, respectively.

6. Conclusion

In this paper, we have presented an approach for integrating multi-modal heterogeneous data sources through a spatial KG layer in the context of integrating geographically annotated microbial data and environmental features to enable joint analysis. While the emphasis has been on integrating the spatial data sources, we have also discussed the design of the complete multi-modal data integration of raster data, DNA reads, and ontologies. We propose an approach based on the S2 Geometry that can integrate raster files of different resolutions, translations, and rotations without performing significant aggregations of the raster cell measurements. In the future, we plan to work on improving scalability given the large number of RDF triples related to the use of the S2 Geometry as well as to capture provenance through the use of the PROV-O ontology¹⁰.

Acknowledgments

This work is partially funded by the Villum Foundation (DarkScience project, 50093). We further thank the SustainScapes group from Aarhus University for sharing the environmental dataset.

¹⁰<https://www.w3.org/TR/prov-o/>

References

- [1] M. Cheatham, C. Pesquita, Semantic Data Integration, in: Handbook of Big Data Technologies, Springer, 2017, pp. 263–305.
- [2] R. Stahl, P. Staab, Why Is Data Integration So Hard?, in: Measuring the Data Universe, Springer, 2018, pp. 23–34.
- [3] J. H. Phan, et al., Integration of multi-modal biomedical data to predict cancer grade and patient survival, in: BHI, IEEE, 2016, pp. 577–580.
- [4] X. L. Dong, D. Srivastava, Big Data Integration, Proc. VLDB Endow. 6 (2013) 1188–1189.
- [5] K. Timmis, et al., The contribution of microbial biotechnology to sustainable development goals, Microbial biotechnology 10 (2017) 984–987.
- [6] V. Bellandi, et al., Toward a General Framework for Multimodal Big Data Analysis, Big Data 10 (2022) 408–424.
- [7] X. Zhu, et al., Multi-Modal Knowledge Graph Construction and Application: A Survey, IEEE Trans. Knowl. Data Eng. 36 (2024) 715–735.
- [8] M. Leida, A. Gusmini, J. Davies, Semantics-aware data integration for heterogeneous data sources, J. Ambient Intell. Humaniz. Comput. 4 (2013) 471–491.
- [9] J. J. Assmann, et al., EcoDes-DK15: High-resolution ecological descriptors of vegetation and terrain derived from Denmark’s national airborne laser scanning data set, ESSD 14 (2022) 823–844.
- [10] L. C. o. Gomes, Soil assessment in Denmark: Towards soil functional mapping and beyond, Frontiers in Soil Science (2023).
- [11] A. Ghosh, M. Simkus, D. Calvanese, Semantic Querying of Integrated Raster and Relational Data: A Virtual Knowledge Graph Approach, in: RuleML+RR (Companion), 2023.
- [12] K. Kyzirakos, et al., GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings, J. Web Semant. 52-53 (2018) 16–32.
- [13] N. Gür, et al., A foundation for spatial data warehouses on the Semantic Web, Semantic Web 9 (2018) 557–587.
- [14] A. B. Andersen, et al., Publishing Danish Agricultural Government Data as Semantic Web Data, in: JIST, 2014, pp. 178–186.
- [15] K. Patroumpas, et al., TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples, in: EDBT/ICDT Workshops, volume 1133 of CEUR Workshop Proceedings, CEUR-WS.org, 2014, pp. 275–278.
- [16] B. Tran, et al., Semantic Integration of Raster Data for Earth Observation: An RDF Dataset of Territorial Unit Versions with their Land Cover, ISPRS Int. J. Geo Inf. 9 (2020) 503.
- [17] R. Zhu, et al., Environmental Observations in Knowledge Graphs, in: DaMaLOS, 2021, pp. 1–11.
- [18] K. Janowicz, et al., Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence, AI Mag. 43 (2022) 30–39.
- [19] C. Shimizu, et al., A Pattern for Features on a Hierarchical Spatial Grid, in: IJCKG, ACM, 2021, pp. 108–114.
- [20] P. L. Buttigieg, et al., The environment ontology: contextualising biological and biomedical entities, J. Biomed. Semant. 4 (2013) 43.