Know, Know Where, KnowWhereGraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence

Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby K. Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Lopez Carr, Andrew Schroeder, Dave Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio and Zhining Gu

Knowledge graphs are a novel paradigm for the representation, retrieval, and integration of data from highly heterogeneous sources. Within just a few years, knowledge graphs and their supporting technologies have become a core component of modern search engines, intelligent personal assistants, business intelligence, and so on. Interestingly, despite large-scale data availability, they have yet to be as successful in the realm of environmental data and environmental intelligence. In this paper, we will explain why spatial data requires special treatment, and how and when to semantically lift environmental data to a knowledge graph. We will present our KnowWhereGraph that contains a wide range of integrated datasets at the human-environment interface, introduce our application areas, and discuss geospatial enrichment services on top of our graph. Jointly, the graph and services will provide answers to questions such as 'what is here', 'what happened here before', and 'how does this region compare to . . . ' for any region on earth within seconds.

Key words: Knowledge Graphs, Semantic Web, Ontology, Interoperability

Introduction and Motivation

Successful decision-makers have strong situational awareness. They have a comprehensive understanding of the context in which their actions will play out. In our global, fast-paced, and densely interconnected world, this context stems from a wide range of heterogeneous resources that span the physical and social sciences. For instance, decision-makers at humanitarian relief organizations need an immediate understanding of physical perils and the regions they affect. When a hurricane causes a disaster, getting supplies to the local population at the right time and location is key. Relief coordinators also need information about previous events such as cholera outbreaks that may have affected the region before the hurricane makes landfall and

Janowicz, Schildhauer, Zhu, Cai, Mai, Mecum, Liu, Shi: University of California, Santa Barbara; Hitzler, Shimizu, Zalewski, Zhou: Kansas State University; Li, Wang, Tian, Gu: Arizona State University; Rehberger, Gonzalez, D'Onofrio: Michigan State University; Fisher, Smith: Hydronos Labs and Oliver Wyman; Lopez Carr, Schroeder: Direct Relief; Wright: Esri

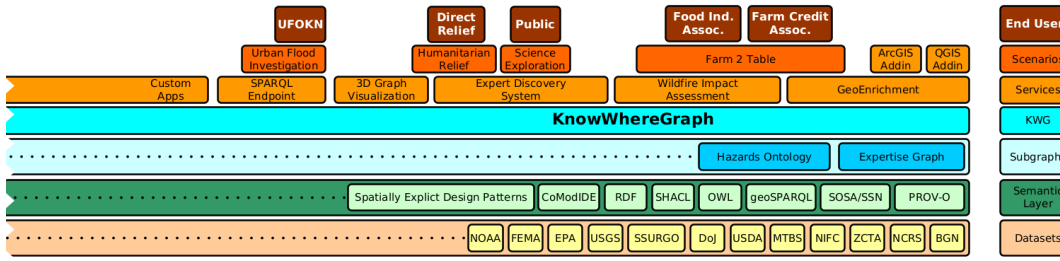


Figure 1. : A layer-wise depiction of the architecture of KnowWhereGraph and the services and use-cases that it supports (as of August 2021).

experts on the ground who can coordinate relief.

Similarly, the agricultural sector, including government agencies, food industry associations, individual farmers, and retailers, requires immediate access to data about food safety, wildfires, floods, air pollution, worker health, supply chain disruptions, and transportation networks. For instance, our partners at the Food Industry Association (FMI) want to understand how a wildfire at one place may impact leafy greens, grapes, and the health of workers at another place 100 miles away due to heavy smoke and ashes. Making decisions based on such data is called *Environmental Intelligence* and is gaining traction due to increased environmental stress, correlated shocks, just-in-time supply chains, and a growing interest in *Environmental*, *Social*, and *Corporate governance* (ESG).

Unfortunately, for practical data-driven decision-making and data science, the first stages of gaining situational awareness consume 80% of a project's resources, be it funds, time, or person power. This leaves merely 20% of the resources for the actual analysis that determines the quality of the decisions. More concretely, most resources are spent on data retrieval, cleaning, and integration rather than on deriving insights from data. This puts data-driven decision-making out of range for many tasks. Several solutions to this well-known *data acquisition bottleneck* have been proposed, both in industry and academia. Most either target the retrieval problem by envisioning one-stop data portals or aim at cloud-based access and processing of data.

In the realm of Geographic Information Systems (GIS), one partial solution are geo-enrichment services. For instance, Esri's GeoEnrichment service enables analysts to enrich their local data on-demand with a range of up-to-date demographic variables apportioned to their area of concern and need. This has a number of advantages: (1) In theory, data is always up-to-date and does not age on the analyst's hard disk; (2) In times of misinformation and information overload, the data comes from a trusted resource; (3) the data is tailored (apportioned) to the analyst's study area; and finally, (4) the data is GIS-ready in the sense that it can be directly processed, analyzed, and displayed. While current geo-enrichment services are valuable, they also face four key limitations: (1) They only serve data for a small set of predefined categories, such as demographic data. (2) They are closed data silos that encode just one domain/cultural perspective. (3) Because they are centrally maintained, scalability and timely updates become bottlenecks when those services try to incorporate more (diverse) data. (4) They do not have an integration layer that enables follow-up queries over the enriched data. Consequently, a new approach is needed that combines the strength of geo-enrichment services, i.e., seamless access to contextual information for an analyst's areas of concern, with a technology that provides open, densely integrated, cross-domain data across a wide range of perspectives (Janowicz (2021)).

For these challenges, knowledge graphs (KGs) promise to provide a solution (Noy et al. (2019); Hogan et al. (2020)). They are a combination of technologies, specifications, and data cultures for densely interconnecting (Web-scale) data across domains in a human and machine readable and reasonable way. They are a novel approach to publishing,

representing, integrating, and interlinking individual data (not merely datasets) by concentrating on connections among places, people, events, and entities instead of their properties. More formally, a KG (as a set of node-edge-node statements called triples) can be thought of as a node and edge labeled directed multigraph. While the term knowledge graph itself does not prescribe any particular technology stack, the largest publicly available KG is the Linked Data cloud based on the RDF/Semantic Web technology stack (Bizer, Heath, and Berners-Lee (2011)). Interconnected statements can be of the form ThomasFire → affected → SantaBarbara and SantaBarbara → partOf → California. Together with schemata (ontologies) specified in knowledge representation languages, these triples would entail a third triple, namely that the Thomas Fire happened in California. As these ontologies encode the semantics of the used terminology, they foster interoperability without restricting semantic heterogeneity (Janowicz et al. (2015); Hitzler (2021)).

Inspired by open knowledge graphs such as DBpedia (Lehmann et al. (2015)) and Wikidata (Vrandečić and Krötzsch (2014)) and services such as GeoEnrichment, our KnowWhereGraph provides a densely connected, cross-domain knowledge graph and geo-enrichment services for a wide range of applications in environmental intelligence by giving decision-makers and data analysts on-demand access to area briefings at a high spatial and temporal resolution for any location on the surface of the earth. To do so, we translate data about extreme events, administrative boundaries, soils, crops, climate, transportation, and so on, into a KG and pre-integrate them to provide answers to questions such as 'what is here', 'what happened here before', 'how does this region compare to . . . '. While DBpedia and Wikidata contain only rudimentary information about places/regions, such as their populations, we give rapid access to information such as the wildfires that have affected an area, the major transportation axis crossing a certain region, and the type of crops and soils present in a given region.

Technological Approach

KnowWhereGraph is quickly and continuously growing as new data silos are identified, and subsequently integrated into our graph, based on the needs of our users and application scenarios. We have developed a number of techniques and ontologies to aid in growing and maintaining KnowWhereGraph. Figure 1 shows a layer-wise view of KnowWhereGraph, as well as the services and use-cases it supports, which directly correspond to many of our techniques.

First and foremost, many of our data sources naturally overlap in space and time and we need to manage a vast amount of heterogeneous spatial data. To do so, we partially depart from traditional linked data approaches that often represent spatial regions as points or polygons on the earth's surface. Instead, we utilize a Discrete Global Grid (Bondaruk, Roberts, and Robertson (2020)) called the ''S2 Grid System''. This lays a hierarchical grid over the earth's surface; each grid cell in a level is comprised of four subcells of increasing spatial resolution. KnowWhereGraph serves data at least at S2 Level 11 (about 20 km^2 per cell) for the USA. However, some regions may have a substantially higher resolution based on data availability, rates of change, and application needs. This approach provides a compromise between data precision and access speed in such a way that it does not preempt downstream, finer-grained topological investigations of the original geometries. Figure 2 depicts selected triples from KnowWhereGraph about regions affected by a hurricane, the impacts, and experts on storm-related topics. In addition to grid cells, we serve many other region identifiers with globally unique IDs so that users can request information about them or interlink and thereby enrich their own data. Examples include, FIPS codes, ZIP codes, media market areas, national weather zones, administrative areas, and gazetteer features, and so on.

Using the S2 grid system as a base, we developed a design pattern[1] for easily relating how features and regions may interact throughout the hierarchy. Additionally, we have adopted a number of open standards such as *GeoSPARQL*[2] and the *Sensors, Observations, Sample, Actuator (SOSA)* ontology[3] and its extension (Zhu et al. (2021)), as well as other frequently used ontologies

[1] https://github.com/KnowWhereGraph/hierarchical-cell-features
[2] https://www.ogc.org/standards/geosparq
[3] https://www.w3.org/TR/vocab-ssn/

Figure 2. : Triples from KnowWhereGraph about a hurricane, impacted areas, impacts, and experts on relevant topics.

such as *QUDT*[4]. Modelling all data from a sensor & observation perspective eases querying, connecting data to the geographic features they describe, and also enables us to link data about events with human experts and research results. Finally, we also worked on connectivity and coverage of our graph. In particular, we provide

- enriched representations of regions, such as climate divisions or counties, and link them to entities from Wikidata or the Geographic Names Information System, where possible, giving instant access to a wide range of broad contextual information such as population density, previous extreme events, soil health, and so on;
- topological relations (e.g., RCC8) among regions for flexible inference and triple compression; and
- link together events and places through causal relationships and provenance (Shimizu et al. (2021)). For instance, we model where a fire took place, which events it triggered, and which regions have been affected, e.g., by heavy smoke.

Altogether, this allows domain scientists to represent geospatial objects, which are traditionally represented as vector geometries, as a collection of S2 cells at various hierarchical grid levels and instantly have tight integration with any other dataset in KnowWhereGraph. In Figure 2, for example, we focus on named places such as counties, but users may request storm damage for any collection of S2 grid cells. The level of S2 cells is not uniform across regions but depends on data layers and (in the future) also on the underlying variation within these layers across space.

Challenges and Relation to Artificial Intelligence

Knowledge graph technology is to a substantial part based on Knowledge Representation (KR) methods and thus on the corresponding subfield of Artificial Intelligence (Hitzler (2021)). In particular, the central W3C standards RDF (Resource Description Framework Cyganiak, Wood, and Lanthaler (2014)) and OWL (Web Ontology Language (Hitzler et al. (2012))) for representing graphs and their schemas (known as ontologies), are formal logics in the tradition of the KR field (Hitzler, Krötzsch, and Rudolph (2010)).

However, in contrast to traditional lines of KR research, recent developments in knowledge graph data management shift the focus to pragmatics, in particular how to make knowledge representation work in practice--at industrial scale, functionality, and stability levels--for data management. While

[4] http://www.qudt.org/

traditional academic literature on KR has a heavy focus on developing KR languages and provably correct and theoretically analyzed algorithms, pragmatic aspects such as the question which KR approach works the best in which situation, or how to apply a KR framework or representation language to an industry scale problem, have played a minor role in academic outlets. Similar questions such as how to lift individual data to the graph, when to do so, at which resolution (e.g., level of granularity), and how to balance schema complexity between optimizing for use versus re-use remain largely unanswered.

KnowWhereGraph focuses on this transition gap between theoretical results and applicability in practice. In particular, it is about the general question of how to achieve practically relevant levels of scale and speed based on real high-volume heterogeneous data from diverse sources, and how to do this without an undue compromise of the quality in representation and solutions that come out of the KR field. In other words, KnowWhereGraph is about finding the right trade-off between principled approaches and rapid, scalable development. It is about finding the sweet spot between theory and practice. In case of KnowWhereGraph, this happens in the context of a multidisciplinary setting that requires rapid convergence across topics such as climate forecasts, extreme events, health, supply chains, and even the spatio-temporal bounds of human expertise for our pilot in disaster relief. Integrating these datasets also requires solutions that can handle noisy and missing (and contradictory) data, as well as changes in perspective as they relate to different schema, and services that enable data exploration using similarity-based search. To handle real-world and noisy data, our work combines symbolic and sub-symbolic methods for representation and reasoning.

Particular challenges related to Artificial Intelligence that we address are (1) bringing principled KG methods to a level of maturity sufficient for transfer to industrial practice, (2) scaling up of methods and processes for our applications for which we currently project a required knowledge graph containing about 10 billion triples, and (3) knowledge graph methods and tools development that is aimed at maximum flexibility for future growth, extension, and reuse.

Specific innovations within the KnowWhereGraph work that are relevant for Artificial Intelligence include:

- In terms of representation of spatial knowledge, we have combined hierarchical grids with standard region boundaries and Region Connection Calculus methods (Zalewski, Hitzler, and Janowicz (2021)), in what we believe is a novel approach for knowledge graphs to meet scale and uniformity requirements.
- In terms of access to large-scale spatial data, we are integrating knowledge graph and GIS technology by offering graph-based geo-enrichment and n-degree property path queries from within a GIS.
- With respect to knowledge representation methods, we are combining top-down and bottom-up ontology engineering processes with a principled modular approach to knowledge graph schema development to balance between quality of the graph model and speed of development and integration.

Current Status

While KnowWhereGraph can serve a wide range of domains and use cases that require spatial data and spatial question answering, we have three initial pilots:

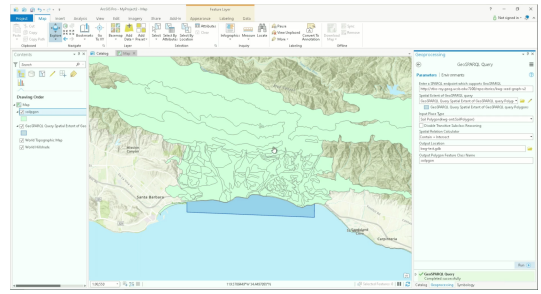
- Humanitarian relief: Together with Direct Relief we demonstrate how our technologies can inform humanitarian supply chains and help identify and match domain experts to the needs of an emerging crisis.
- Farm to Table Supply Chain & Sustainability: In collaboration with the Food Industry Association, we demonstrate how knowledge graphs can enhance the sustainability, efficiency, and safety of consumer food supply with a focus on the impact of wildfires on agriculture and food security.
- Land Valuation and Risk of Default: This new pilot is a joint research with farm credit associations concerned with driver-based land potential assessment for model based valuation and risk assessment for agricultural credit applications and loan portfolio monitoring.

To date we have included 27 different data layers from 16 major data sources that extensively cover the topics discussed in the domain application areas (e.g., climate hazard, wildfire, and air quality). At the time of

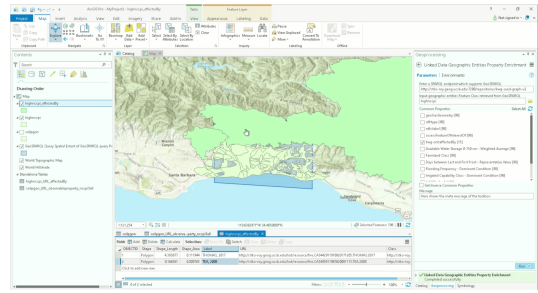
writing, KnowWhereGraph already consists of about 4.9B triples, and we expect it to grow to as many as 10-20B triples over the next years as we ingest additional data. We will also provide area briefings at even higher S2 cell resolution, achieve global coverage (beyond our mostly US-centric data), as well as mine new and more complex relationships across the described places and events.

Built upon the KnowWhereGraph, our geo-enrichment services provide a set of toolboxes which support domain scientists to explore environment-related knowledge from within a GIS in various ways such as region-based spatial data retrieval (e.g., soil polygons can be retrieved based on a user-defined study area as shown in Figure 3a), property enrichment for geographic entities (e.g., querying a crop productivity index for each of millions of soil polygons), direct relation exploration among geographic entities (e.g., querying for landslides on soils previously affected by wildfires as shown in Figure 3b), and n-degree relation identification (e.g., SoilPolygonA → affectedBy → ThomasFire → causedEvent DebirsFlowX).

We have also developed a range of additional services tailored to our vertical applications. For example, the KnowWhereGraph enables disaster relief specialists to explore knowledge about experts and their areas of expertise, as related to specific disasters. To achieve this, we provide a similarity search interface and a follow-your-nose interface, which are shown in Figure 4. In case of the similarity interface, users can type in an expert name into the search box and the system will return the top 15 experts who are most similar. The similarity score is computed using a combination of Doc2Vec and knowledge graph embedding techniques (Le and Mikolov (2014); Mai, Janowicz, and Yan (2018)), which are computed based on the particular expert's three most cited papers, three most recent papers, and their relation to other experts in the graph. Figure 4 (left) shows an example of the similarity search. From there, users can directly search information about the experts, their area of expertise, and events that they have worked on. Conversely, users can start by selecting a certain event or a geographic region, learn about previous events, their impacts, and the relevant experts that could be contacted. In fact, this ability to seamlessly navigate between physical events, areas of expertise, affected regions, and people is one of the key strengths of our knowledge graph.



(a) Retrieving soil polygons



(b) Retrieve wildfires that affected soil polygons

Figure 3. : Our Knowledge Graph b ased geo-enrichment toolbox collections for ArcGIS Pro. (a) The GeoSPARQL Query toolbox (b) The Property Enrichment toolbox.

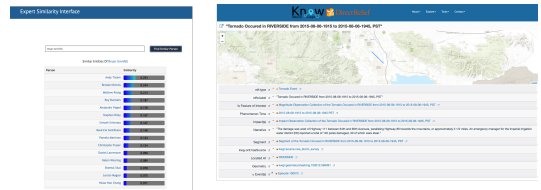


Figure 4. : Left: Similarity interface for experts. Right: Follow-your-nose interface for previous disasters.

In terms of our food safety work, KnowWhereGraph is used to enhance assessment and strategic planning during near real-time hazard events affecting the food supply chain by providing online analysis, forecasting, and alerts that are enriched with location and context-specific intelligence,

to ensure key stakeholders throughout the supply chain are ready with backup strategies to keep products moving. It also allows farmers and growers to identify how they can be better prepared to mitigate and build resilience in the face of such events. Currently, our graph serves pre-integrated data about wildfires, smoke plumes, and crop locations, together with topological information about the affected areas. In one implementation for FMI, a custom front-end web interface (Figure 5) and API enables decision-makers to process a series of queries important to assessing the impact of ongoing wildfires, smoke plumes, and ashes on key food (crop) supply chains. Users can progress through these queries without any experience in using complex GIS software or the specific data and analysis techniques necessary, seeing visualizations of the results at each step. Despite the simplicity of this system, the interface is dynamically generating SPARQL queries based on the user inputs (e.g., defining a region of interest, selecting multiple crop types), sending these queries to the graph via an API and receiving/displaying the results, all within a matter of seconds. This system highlights the ease with which new bespoke end-user applications can be developed from the core resources of the KnowWhereGraph, enabling a multitude of use cases at the human-environment interface.

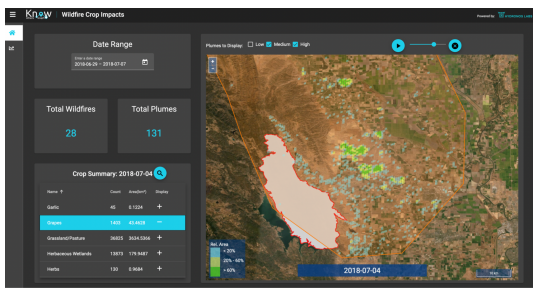


Figure 5. : Wildfire crop impacts interface displaying a smoke plume (yellow shape) from July 4, 2018 associated with the County Fire (red outlined shape). Within this plume we have queried for areas with high densities of grapes to identify areas where the crop may be affected by smoke taint.

Future Plans

In this work, we have introduced the KnowWhereGraph, a densely connected, cross-domain knowledge graph together with geo-enrichment services to support a variety of application areas that benefit from environmental intelligence. Our graph delivers area briefings for any place on earth within seconds to answer questions such as 'what is here' or 'what happened here before'. For instance, decision-makers and data scientists can easily retrieve all extreme events (e.g., previous storms, fires, cholera outbreaks) that have impacted an area that is predicted to be in the path of an approaching hurricane. Most importantly, we do not only serve data layers, but also connections across them. For instance, graph hubs such as Wikidata or DBpedia contain information about Santa Barbara, the Thomas Fire, highway 101, and the 2018 debris flow in Southern California. However, they do not locate the fire nor the debris flow and most importantly do not contain facts such that the fire affected Santa Barbara and that the fire and a massive storm caused a debris flow that killed 23 people and disrupted transportation for weeks as it blocked highway 101. This is exactly the type of relationships that we are most interested in. We also do not just serve data at predefined levels, e.g., counties, but deliver a variety of regions identifiers thereby making KWG a gazetteer of gazetteers. I addition, we also serve data registered to fine-grained global grid cells. So far, the KnowWhereGraph largely contains information about the US due to easy access to high-quality, well-documented governmental data, as many of our use cases revolve around the US, and to keep the graph size at bay. In the future, we will increase global coverage, add more data layers, enable geo-enrichment for open source GIS and spatial statistics packages in general, and mine more (complex) relationship across our entities with the ultimate goal of creating a global knowledge graph of environmental and geographic information.

Acknowledgments

The authors acknowledge support by the National Science Foundation under Grant 2033521 A1: KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs

using Spatially-Explicit AI Technologies. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bizer, C.; Heath, T.; and Berners-Lee, T. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*. IGI global. 205--227.

Bondaruk, B.; Roberts, S. A.; and Robertson, C. 2020. Assessing the state of the art in discrete global grid systems: Ogc criteria and present functionality. *Geomatica* 74(1):9--30.

Cyganiak, R.; Wood, D.; and Lanthaler, M., eds. 2014. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation 25 February 2014. Available from http://www.w3.org/TR/rdf11-concepts/.

Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P. F.; and Rudolph, S., eds. 2012. *OWL 2 Web Ontology Language: Primer (Second Edition)*. W3C Recommendation 11 December 2012. Available from http://www.w3.org/TR/owl2-primer/.

Hitzler, P.; Krötzsch, M.; and Rudolph, S. 2010. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press.

Hitzler, P. 2021. A review of the semantic web field. *Commun. ACM* 64(2):76--83.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Gayo, J. E. L.; Kirrane, S.; Neumaier, S.; Polleres, A.; et al. 2020. Knowledge graphs. *arXiv preprint arXiv:2003.02320*.

Janowicz, K.; Van Harmelen, F.; Hendler, J. A.; and Hitzler, P. 2015. Why the data train needs semantic rails. *AI Magazine* 36(1):5--14.

Janowicz, K. 2021. Knowwheregraph drives analytics and cross-domain knowledge. *ArcUser* 16--19.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188--1196. PMLR.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; and Bizer, C. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6(2):167--195.

Mai, G.; Janowicz, K.; and Yan, B. 2018. Combining text embedding and knowledge graph embedding techniques for academic search engines. In *Semdeep/NLIWoD@ ISWC*, 77--88.

Noy, N. F.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; and Taylor, J. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62(8):36--43.

Shimizu, C.; Zhu, R.; Schildhauer, M.; Janowicz, K.; and Hitzler, P. 2021. A pattern for modeling causal relations between events. In *Proceedings of the 12th Workshop on Ontology Design and Patterns (WOP 2021), co-located with the 20th International Semantic Web Conference (ISWC 2021) : online, October 24, 2021*, volume 3011, 38--50.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78--85.

Zalewski, J.; Hitzler, P.; and Janowicz, K. 2021. Semantic compression with region calculi in nested hierarchical grids. In Meng, X.; Wang, F.; Lu, C.; Huang, Y.; Shekhar, S.; and Xie, X., eds., *SIGSPATIAL '21: 29th International Conference on Advances in Geographic Information Systems, Virtual Event / Beijing, China, November 2-5, 2021*, 305--308. ACM.

Zhu, R.; Ambrose, S.; Zhou, L.; Shimizu, C.; Cai, L.; Mai, G.; Janowicz, K.; Hitzler, P.; and Schildhauer, M. 2021. Environmental observations in knowledge graphs. In *2nd Workshop on Data and research objects management for Linked Open Science*.