Received: 14 May 2021 | Accepted: 5 November 2021

DOI: 10.1002/aaai.12035

SPECIAL TOPIC ARTICLE

Knowledge graphs to support real-time flood impact evaluation

J. Michael Johnson[1] | **Tom Narock**[2] | **Justin Singh-Mohudpur**[1] | **Doug Fils**[3] | **Keith C. Clarke**[1] | **Siddharth Saksena**[4] | **Adam Shepherd**[5] | **Sankar Arumugam**[6] | **Lilit Yeghiazarian**[7]

[1]University of California, Santa Barbara

[2]Goucher College

[3]Consortium for Ocean Leadership

[4]Virginia Polytechnic Institute and State University

[5]Woods Hole Oceanographic Institution

[6]North Carolina State University

[7]University of Cincinnati

Correspondence
Tom Narock, Goucher College, 1021 Dulaney Valley Rd, Baltimore, MD 21204.
Email: Thomas.Narock@goucher.edu

Funding information
NSF, Grant/Award Numbers: OIA #1937099, #2033607

Abstract

A digital map of the built environment is useful for a range of economic, emergency response, and urban planning exercises such as helping find places in app driven interfaces, helping emergency managers know what locations might be impacted by a flood or fire, and helping city planners proactively identify vulnerabilities and plan for how a city is growing. Since its inception in 2004, OpenStreetMap (OSM) sets the benchmark for open geospatial data and has become a key player in the public, research, and corporate realms. Following the foundations laid by OSM, several open geospatial products describing the built environment have blossomed including the Microsoft USA building footprint layer and the OpenAddress project. Each of these products use different data collection methods ranging from public contributions to artificial intelligence, and if taken together, could provide a comprehensive description of the built environment. Yet, these projects are still siloed, and their variety makes integration and interoperability a major challenge. Here, we document an approach for merging data from these three major open building datasets and outline a workflow that is scalable to the continental United States (CONUS). We show how the results can be structured as a knowledge graph over which machine learning models are built. These models can help propagate and complete unknown quantities that can then be leveraged in disaster management.

INTRODUCTION

As cities become more complex, integrative, and digital, there is a need to understand where urban features are, and what is their primary purpose. Recognizing the interconnectedness of systems such as buildings, transportation, power, and water is critical to the success, safety, and sustainability of urban regions. Collectively, this system of systems is referred to as an Urban Multiplex. Currently, no information system exists where all of this infrastructural

information is collected and aggregated. We are developing a first of its kind Urban Multiplex Inventory (UrMI) using knowledge graph technologies.

Within any urban multiplex, buildings serve as a lynchpin determining where many interactions occur, and how these interactions impact human well-being, comfort, and at times, life. Examples of these interactions are reflected in a range of queries ranging from "how do we get a burrito delivered from this restaurant to this home" to "which buildings to evacuate during a flood."

Yet, there is a major discrepancy in the way building and related multiplex data is collected across the United States. There is no harmonization in what data are recorded by city, county, or state governments let alone at a national scale. Most previous efforts to harmonize the urban multiplex are proprietary, owned by groups such as Google, Yelp, and Zillow, and utilized solely in their respective apps. Our goal is to have a uniform, relatively open dataset that captures the critical information about urban infrastructure. To this end, there are three primary open datasets of urban infrastructure, which form the foundation of our project.

However, these three open datasets are driven by different organizational priorities and inclusion criteria. OpenStreetMap (OSM) features provide verifiable building footprints along with key and value annotations. Microsoft building footprints (Microsoft, 2018) are built from a machine-learning approach to provide an unprecedented level of completeness; and the OpenAddress (OA) project aggregates address information from authoritative agencies but provides no information on the building area, type, or purpose. For many urban-environmental analyses, there is a critical need for information about location (all), area (MS), address (OA), and building type (OSM) but working between and across these datasets can be a challenging task. Not only do they come with different information, but they have no shared identifier outside of spatial location. Similarly, they are represented in a variety of data formats including pbf, geojson, shp, and CSV. While there is an opportunity to integrate these datasets under a single resource, there is an equal opportunity to structure them in a way that increases interoperability and their use outside of more traditional GIS environments to support advanced analytics and additional data enhancements.

One way to enhance the interoperability of these datasets is using knowledge graphs. Although the term *knowledge graph* was popularized with the release of Google's Knowledge Graph in 2012 (Hitzler 2021), the definition of "knowledge graph" remains contentious (see Hogan et al. 2021 and references therein). Here, we adopt the inclusive Hogan et al. (2021) definition that a *knowledge graph* is "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities." Much of the contention in defining knowledge graphs comes from the breadth and depth of their design and use. Knowledge graphs can be openly accessible as in the case of Wikidata (Vrandečić and Krötzsch 2014), or they can be enterprise knowledge graphs (e.g., efforts by Google, Facebook, and eBay) serving internal business needs and not entirely accessible outside of the company (Noy et al. 2019; Hitzler 2021). Additionally, data can be structured as a graph in several different ways with some of the more popular techniques including *directed edge-labeled graphs* and *property graphs* (Angles et al. 2017; Hogan et al. 2021). Placed in this context, UrMI is a collection of both open and enterprise directed edge-labeled knowledge graphs encoded using the standardized data models Resource Description Framework (RDF, Cyganiak, Wood, and Lanthaler 2014) and Web Ontology Language (OWL, Hitzler et al. 2009).

TECHNICAL APPROACH

It is often desirable to manage several graphs rather than one monolithic graph (Hogan et al. 2021). This approach makes it possible to update or refine data independently and in parallel. UrMI is built on three knowledge graphs with the capability to add more use case focused graphs for additional environmental shocks (i.e., floods or wildfires) or contextual data. The first graph describes the building resources – hereby UrMIStructure(s) – while the second divides the US into geographical regions that may be queried individually or collectively. The UrMIStructure graph pattern (see Figure 1) is instantiated with initial data from OSM. Where available, the Address is instantiated from Open Address information and geometry coming from Open Street Map and/or Microsoft. CollectionOfFeatures allows grouping related features. For example, a sports arena and its parking garage are each described as an UrMIStructure. The two instances together form a CollectionOfFeatures instance. A benefit of the knowledge graph's Open-World Assumption is that it accommodates all combinations of OSM-MS-OA availability. For example, buildings are known to have addresses even though we may not know them for every feature (e.g., MS or OSM-MS UrMI structures). The third UrMI knowledge graph is a socio-economic graph built from the U.S. Census' American Community Survey (ACS). The ACS graph enhances risk assessment due to environmental shocks with additional demographic and socio-economic context.

It is anticipated that the UrMI knowledge graph triple count will be in the billions when UrMI graph patterns are instantiated for the continental United States (CONUS). This presents a daunting challenge for graph maintenance and efficient query execution. These challenges are muted

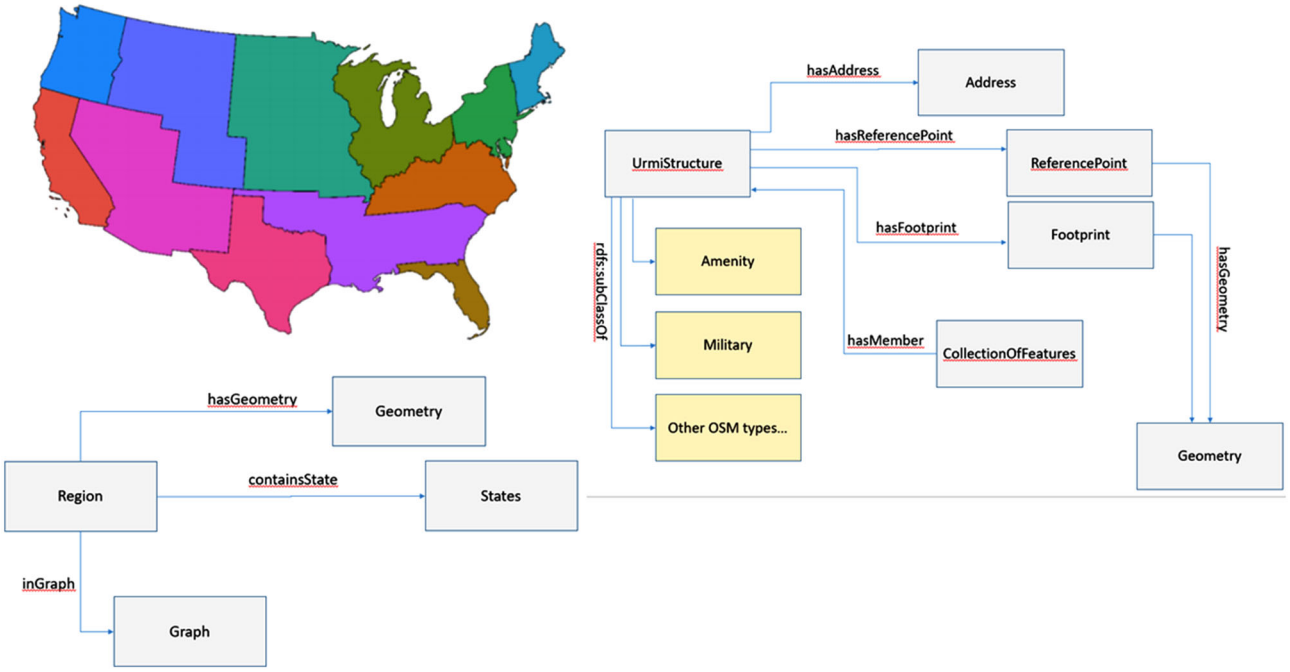

FIGURE 1 The CONUS is divided into twelve areas to manage graph size and query execution. Examples of the UrMI structure graph and the CONUS Regions graph are also shown. CONUS, continental United States; UrMI, Urban Multiplex Inventory

by the second knowledge graph – the CONUS regions graph (Figure 1). This second graph divides the CONUS into 12 regions, each of which has its own named graph in the UrMI graph database. Each of the twelve regions has a Geometry (the polygon outlining the region), a list of states included in that region, and a reference to the named graph containing data from that region. Users (directly via SPARQL queries or from within UrMI apps) can query the region graph with state(s) or county(s) names or a bounding box of interest. The user is returned the relevant graph(s) to query for feature information.

UrMIStructures are populated via an efficient software pipeline that executes data downloading and preprocessing, spatial integration, and RDF/OWL encoding for any county in the US. The pipeline is automated through the use of an Amazon Web Services Lambda service that watches for new or modified county level files.

AUTOMATED INFERENCE AND MACHINE LEARNING

Identifying unknown structures is critical for assessing the number of people impacted by a natural disaster. Yet, of our three source datasets, OSM is the only one that seeks to describe building types through its key and value tags and there is a volunteer preference for tagging buildings in large urban areas. In New Hanover, North Carolina, for example, 22,489 of the county's 25,394 OSM buildings (89 percent) do not have a building type. This is where basic RDF/OWL inferencing becomes beneficial. As an example, consider the case of 105 Tennessee Avenue in New Hanover, North Carolina, which has no OSM associated name/value data. Yet, from OA integration, we find that this building has 22 addresses (105 Tennessee Avenue Unit 100, 105 Tennessee Avenue Unit 101, etc.). This building at 105 Tennessee Avenue is inferred to be a multiunit building through address cardinality axioms. Additional class membership inferences are made using owl:hasValue. For example, the knowledge graph defines Residential to be a subclass of UrmiStructure and includes in this class all UrmiStructure instances whose UrmiStructure#name property is "residential." This simple inference has immense practical benefit in that semantics are removed from the data generation/preprocessing stage and maintained entirely in the knowledge graph. There is no need for the RDF generation software to care about OSM building types and subclasses, some of which are irrelevant to the goals of UrMI. Instead, all urban features are inserted into the knowledge graph simply as instances of UrmiStructure. The knowledge graph then takes care of sorting and classifying features as needed.

The sparseness of truly informative tags speaks to the challenges of volunteer efforts and varying completeness, but also an opportunity to leverage machine learning to fill in information in the graph. We are developing K-Nearest Neighbors and Decision Tree approaches to infer the most

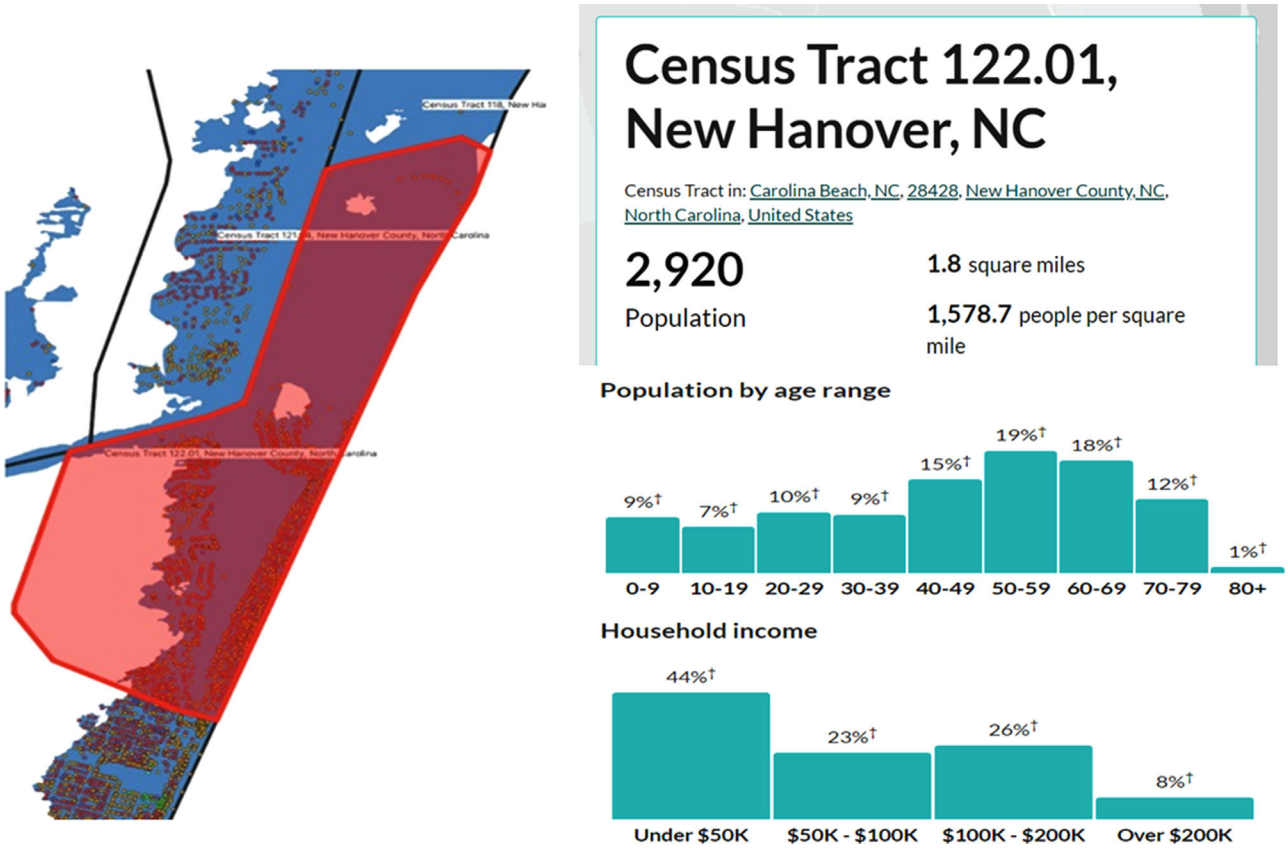

FIGURE 2 Mixing SPARQL, GeoSPARQL, and online Census Data with Urban Multiplex Inventory (UrMI) to estimate flood impacts during the 100-year flood event in New Hanover County, North Carolina

likely type of unknown buildings based on known quantities such as building area and address. In initial tests, we are able to predict that the aforementioned multiunit building is likely an apartment building, which is then inferred to be a type of residential building. The combination of OWL inferences and machine learning takes this entity from being a generic "building" to a large residential structure needing the focus of emergency responders in a flood event.

CHALLENGES AND NEED FOR AI – A FLOOD CASE STUDY

Continuing with the example for New Hanover, North Carolina, we obtained the 100-year flood map from the FEMA National Flood Hazards Layer. The term "100-year flood" is used to describe the extent of a flood that statistically has a 1-percent chance of occurring in any given year (Maidment 2009). Here, it is used for illustrative purposes of an overall disaster response application of UrMI and the need for AI, but other flood representations could come from real-time flood mapping efforts like those being pursued by NOAA (Maidment 2016; Liu et al. 2018; Johnson et al. 2019). Equally, any other disaster that can be represented as a polygon could be used in its place (snow warning, fire extent, etc.).

Flooded UrMIStructures can be identified at the county level using GeoSPARQL *WithIn* queries. The results (Figure 2) indicate that 12,328 New Hanover buildings would be inundated and of these, nearly 60 percent have an associated address and footprint, while 96 percent have an associated address only. Integration with census ACS graph reveals that one tract has an approximate population of 2920 people, of which, 472 are under 19 years old, and 893 are over 60 years old. In addition to flooded buildings, impacts to public transportation have secondary socio-economic impacts. From the census ACS graph, we can learn when most workers are on the roads in this tract (most commute between 7 and 8 am), the general occupational makeup of the tract (management and business), and the vehicles expected on roadways (most inhabitants drive to work alone). While this is a simulated flood event, it demonstrates how a combination of building level structural assessments, paired with neighborhood level socio-economic summaries can

provide a richer context to those seeking to understand a location in the context of the human–environmental interface.

CURRENT STATUS

Here, we have documented an approach for merging data from three major open building datasets and outlined a workflow that is scalable to the CONUS. We have developed three knowledge graphs over which machine learning models are built. These machine learning models help complete unknown quantities that can be leveraged in disaster management. This integrated product can serve as a backbone for linking other data systems, environmental models, and digital representations of the human–environment interface to help answer more integrative and convergent questions in an era of increasing complexity.

At present, UrMI contains data from two US states. Having worked out the software pipeline enabling automated knowledge graph instantiation from OSM, MS, and OA, we are now ready to expand to CONUS scale. As a scalability example, the software pipeline processed the county of New Hanover, North Carolina in about twenty minutes. This led to approximately 1.8 million statements in the knowledge graph with approximately 1.2 million explicitly asserted and about 636,000 inferred. This is an expansion ratio (total/explicit) of 1.52. The 1.8 million statements were inserted into the knowledge graph in just under two minutes once the building file became available, leading to near immediate availability for urban planning and risk assessment.

FUTURE PLANS

The Urban Multiplex sits at the center of many economic, emergency response, and planning exercises. To analyze the Urban Multiplex, there is a need for an UrMI that describes the features comprising the independent systems, as well as the connections between the systems at the feature level. We are developing such an integrated framework at CONUS scale from disparate open datasets that were originally built using different methods, criteria for inclusion, and overarching goals.

The Urban Multiplex information fits naturally into a graph-based data model, and representing the UrMI as a knowledge graph provides an opportunity for answering questions related to system interactions, disasters at the environmental interface, and for connecting multiple representations of relative location. UrMI is developing the infrastructure to do this at scale (CONUS) with plans for adding additional layers and networks embedded in the Urban Multiplex. Future work will document how this process works in production and at scale, how it supports national scale, feature level disaster response, and ultimately, how it can serve as a principal information backbone for our understanding of the human–environment interface in the face of increasing social, environmental, and infrastructural complexity. The adoption of multiple knowledge graphs holds promise for others to leverage, build on, and improve upon this initial release.

ACKNOWLEDGMENT

The authors gratefully acknowledge funding provided by NSF through grants OIA #1937099 and #2033607.

CONFLICT OF INTEREST

The work described in this manuscript is funded by the National Science Foundation (NSF). The NSF encourages, and has enabled, discussion and collaboration with external commercial entities to explore long term viability of our real-time flood impact evaluations. The authors of this work acknowledge such discussions, but certify that, at the time of publication, no author had any affiliation with, or involvement in, an organization or project that could be considered a conflict of interest.

REFERENCES

Angles, Renzo, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan L. Reutter, and Domagoj Vrgoc. 2017. "Foundations of Modern Query Languages for Graph Databases." *ACM Computing Surveys* 50(5):68:1–68:40. https://doi.org/10.1145/3104031

Cyganiak, Richard, David Wood, and Markus Lanthaler. 2014. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. World Wide Web Consortium, February 25, 2014. https://www.w3.org/TR/2014/RECrdf11-concepts-20140225.

Hitzler, Pascal 2021. "A Review of the Semantic Web Field." *Communications of the ACM* 64(2): 76–83.

Hitzler, Pascal, Markus Krötzsch, Bijan Parsia, Peter Patel-Schneider, and Sebastian Rudolph. 2009. "OWL 2 Web Ontology Language Primer." *W3C Recommendation* 27(1): 123.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab & Antoine Zimmermann 2021. Knowledge Graphs, ArXivx, arXiv:2003.0230. https://arxiv.org/abs/2003.02320.

Johnson, J. Michael, Dinuke Munasinghe, Damilola Eyelade, and Sagy Cohen. 2019. "An Integrated Evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) Flood Mapping Methodology." *Natural Hazards and Earth System Sciences* 19(11): 2405–20.

Liu, Yan, David Maidment, David Tarboton, Xing Zheng, and Shaowen Wang. 2018. "A CyberGIS Integration and Computation Framework for High-Resolution Continental-Scale Flood Inundation Mapping." *Journal of the American Water Resources Association* 770–84.

Maidment, David 2016. "Conceptual Framework for the National Flood Interoperability Experiment." *JAWRA Journal of the American Water Resources Association* 53(2): 245–57.
Maidment, David 2009. "FEMA Flood Map Accuracy." *World Environmental and Water Resources Congress 2009: Great Rivers* 1–10.
Microsoft. 2018. *US Building Footprints*.
Noy, Natasha, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. "Industry-Scale Knowledge Graphs: Lessons and Challenges." *Communications of the ACM* 62(8): 36–43.
Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57(10): 78–85.

AUTHOR BIOGRAPHIES

J. Michael Johnson is a recent PhD graduate in Geography from the University of California, Santa Barabara. He is currently working as NOAA Affiliate (Water Resource Engineer II) supporting The National Water Center and their next generation water modeling efforts.

Tom Narock is an Assistant Professor of Data Science and a member of the Center for Data, Mathematical, and Computational Sciences at Goucher College. He has a background in Earth science applications and his research focuses on data science within the geosciences.

Justin Singh-Mohudpur is a scientific software developer affiliated with the University of California, Santa Barbara and the United States Air Force Reserve. Justin develops web and cloud-based hydrological tools as well as machine learning models.

Doug Fils is a Data Management Technical Expert at the Consortium for Ocean Leadership. His responsibilities involve coordinating with other data management staff with the goal of implementing a common architecture for the distribution of ocean and water data resources. Prior to coming to Ocean Leadership, Doug worked with the NSF originated CHRONOS effort establishing services oriented procedures for the distribution of Earth history databases, tools, and services.

Keith C. Clarke is a Professor of Analytical Cartography and Modelling in the Department of Geography at the University of California, Santa Barbara, USA. He is a Fellow of the American Association for the Advancement of Science and his research area is cartography and geographic information science.

Siddharth Saksena is a Research Assistant Professor in the Department of Civil and Environmental Engineering at Virginia Polytechnic Institue and State University. Siddharth is working on developing tools for building flood resilient communities in response to land use and climate change involving watershed-scale flood modeling and forecasting using an integrated hydro-systems analysis.

Adam Shepherd is an Information Systems Associate and Data Manager at Woods Hole Oceanographic Institution. Adam writes scientific software in support of marine data management, discovery and access.

Sankar Arumugam is a Professor in the Department of Civil, Construction, and Environmental Engineering at North Carolina State University, where he is also a University Faculty Scholar (2013-2018). He is primarily associated with the Environmental, Water Resources, and Coastal Engineering and Computing and Systems groups within the department.

Lilit Yeghiazarian is a Professor in the College of Engineering and Applied Science at the University of Cincinnati. Her research focuses on complex systems and processes.

How to cite this article: J. Michael Johnson, Tom Narock, Justin Singh-Mohudpur, Doug Fils, Keith C. Clarke, Siddharth Saksena, Adam Shepherd, Sankar Arumugam, and Lilit Yeghiazarian. 2022. "Knowledge Graphs to support real-time flood impact evaluation." *AI Magazine*. 43: 40–45. https://doi.org/10.1002/aaai.12035