SAWGraph: Safe Agricultural Products and Water Graph

Torsten Hahmann, University of Maine; Onur Apul, Ganga Hettiarachchi, Pascal Hitzler, Hari Prasath Palani

Use case description and societal challenge being addressed

The Safe Agricultural Products and Water Graph (SAWGraph) is an open knowledge network (OKN) for knowledge on contamination of water and food supplies with perfluoroalkyl and polyfluoroalkyl substances (PFAS), and for assessing how different places and communities are exposed to different levels of PFAS-related health risks. Domain knowledge and semantic metadata expressed in KGs add context, bridge data silos, and help infer richer layers of information such as connections between what has been observed and where it may have originated from or what other places or resources it may impact. The graph's development supports three use cases: **(1) Prioritizing Testing and Identifying Testing Gaps; (2) Assessing Impacts and Risk to Different Communities; (3) Connecting Contamination Sources to Results for Improved Tracing.** In addition, the graph will support communication of contamination occurrence and other kinds of hypothesis development to support PFAS research.

The primary end users of SAWGraphs are personnel at state and federal agencies who monitor the environment and the safety of drinking water and agricultural land and products. End users include but are not limited to personnel at EPA and state environmental protection agencies (e.g. state environmental toxicologists), USDA, FDA, USGS and corresponding state agencies. We distinguish four types of users based on their primary responsibility: (1) drinking water monitoring; (2) understanding the extent and sources of environmental contamination with PFAS; (3) PFAS researchers; (4) environmental justice.

The graph is expected to empower these users to: (1) formulate and implement comprehensive test plans aimed at monitoring elevated PFAS levels in drinking water or agricultural lands, (2) pinpoint the most significant sources of PFAS in a state and how it impacts disadvantaged communities; (3) identify populations at elevated risk and prioritize testing resources and support accordingly; (4) design contamination management and remediation plana; and (5) identify research gaps in knowledge about PFAS fate and transport. Collectively, these applications will expedite comprehension of PFAS contamination and its mitigation, while ensuring the efficient and equitable allocation of testing and remediation resources.

Knowledge graph source datasets

1. PFAS testing results from drinking water (federal: PFAS Analytics Tools; original source UCMR data and select state data); ground and surface water, soil and tissue samples.
2. Point sources for PFAS based on the Facilities Registry Service using North American Industry Classification System (NAICS) and Standard Industrial Classification (SIC) as well as DoD sites, DoE sites, airports, and superfund sites and landfills
3. EPA's PFAS release data collected in accordance with the Clean Water and Clean Air Acts
4. Potentially impacted features like private water wells and drinking water protection areas from select states.
5. Data layers on agricultural uses; surface water (stream reaches, watersheds) and groundwater (aquifers) features via connections to geoconnex and KnowWhereGraph.
6. Geospatial relations to connect data: Spatial connection via S2 cells (level 13 or finer) and via administrative regions (geoIds, FIPS codes, ZIP codes, etc.) and hydrologic connections

The graph to be subdivided by state where possible and be constructed only for select states. Each graph would be on the order of 1 million entities and 10 million triples and likely rapidly growing.

User queries / competency queries for the use case

- Testing: Where (locations/towns) should we prioritize environmental/water testing?
  E.g. which wells are within 5 miles of landfills/airports/biosludge application sites? Which wells near locations with a reported PFOA contamination above 4ppt have not been tested?
- Assessing Impacts: Where and who is impacted? (Prevalence, risk/vulnerability mapping)
  E.g. Which towns or counties have multiple test results with PFOS levels above 20ppt with no known contamination source nearby? Which chemicals show the highest average readings in samples from a particular region? Which towns/counties are most vulnerable based on their proximity to known or potential PFAS sources?
- Tracing: What are potential sources of contamination at a particular location or area?
  E.g. What potential contamination sources exist 20 miles upstream from the sample result? What wells are hydrologically connected to other wells with a reported contamination of +10ppt?