OPEN KNOWLEDGE NETWORK ROADMAP:

APPENDIX B

CONVERGENCE ACCELERATOR TRACK A:

OPEN KNOWLEDGE NETWORK PROJECTS

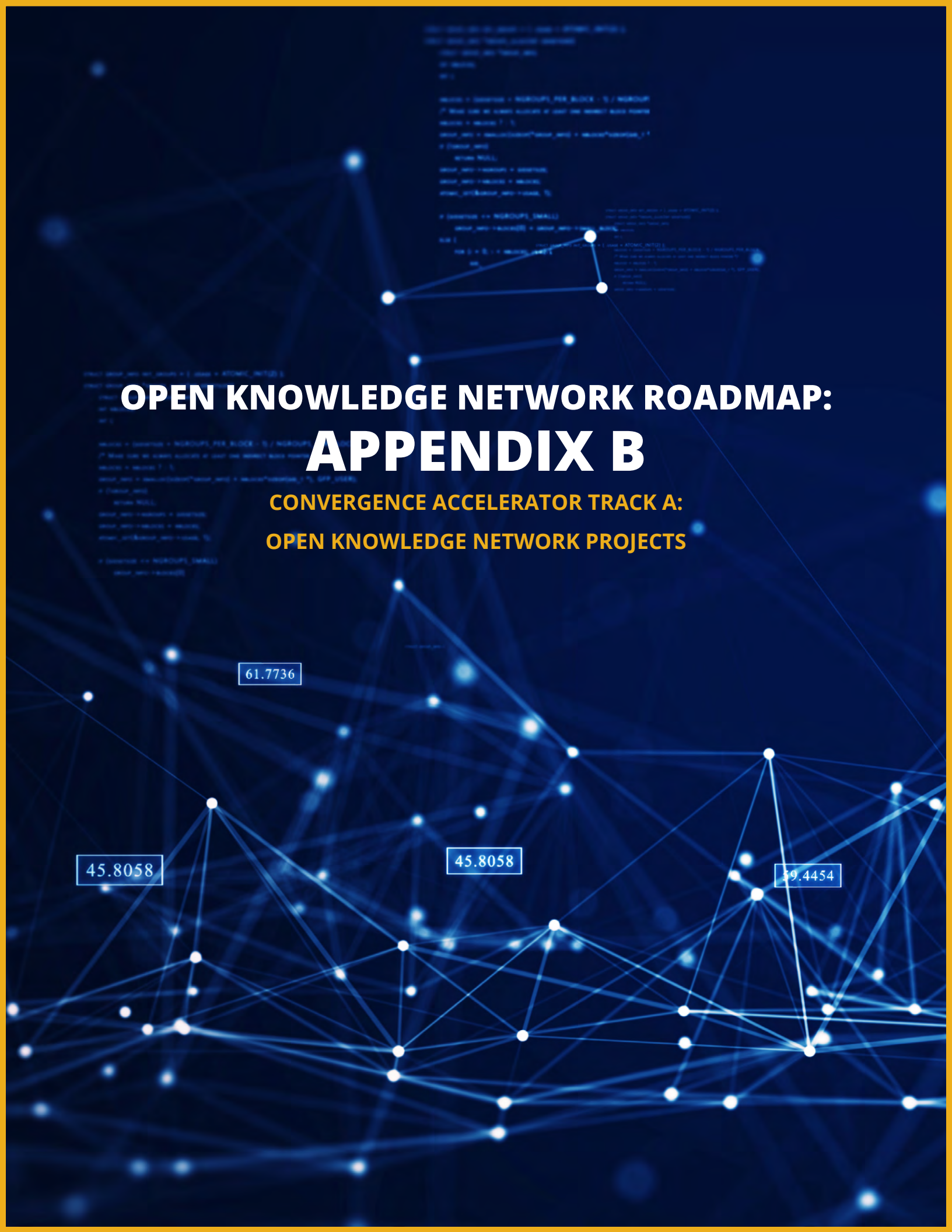


TABLE OF CONTENTS

Convergence Accelerator Track A OKN Projects Overview 4

SCALES - OKN 6

Project goals and accomplishments 6

Videos, websites, and other resources 7

Track integration goals and accomplishments 8

SPOKE 10

Project goals and accomplishments 10

Videos, websites, and other resources 11

Track integration goals and accomplishments 11

Urban Flooding OKN 13

Project goals and accomplishments 13

Videos, websites, and other resources 15

Track integration goals and accomplishments 15

KnowWhereGraph 16

Project goals and accomplishments 16

Videos, websites, and other resources 17

Track integration goals and accomplishments 17

Knowledge Network Programming Infrastructure (KNPI) 20

Project goals and accomplishments 20

Videos, websites, and other resources 22

Track integration goals and accomplishments 22

CONVERGENCE ACCELERATOR TRACK A OKN PROJECTS

Appendix B provides an overview of five Open Knowledge Network (OKN) projects supported by the NSF **Convergence Accelerator Track A Program**. The **Convergence Accelerator** Program builds upon basic research and discovery to accelerate solutions that have substantial societal impact. The program funds teams to address societal challenges through convergence research and innovation. This convergence approach, including a strong focus on human-centered design, brings together the disciplines, approaches, and organizations that are needed to create innovative solutions, refine prototypes, and then implement useful tools and deliverables.

The goal of Track A - Open Knowledge Networks was to enable new modes of data-driven discovery that help advance the progression from data to knowledge. The 21 **Phase 1** projects funded under Track A identified challenges in different topical domains (e.g. flooding, personalized medicine) as well as "horizontal" challenges that cross multiple domains (e.g. incorporating geospatial data, managing provenance). Phase 1 grantees refined their ideas by engaging with stakeholders using human-centered design methods and expanded their multidisciplinary and multi-institutional teams to create effective deliverables.

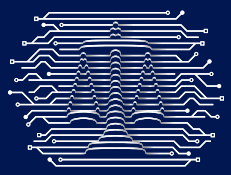
Subsequently, 5 of these 21 projects were selected for a two-year **Phase 2** effort. While user needs were explored in Phase 1, a Minimal Viable Product (MVP) and of a clear set of product deliverables was to be developed in Phase 2. The five Open Knowledge Networks phase 2 efforts include:

- **Systematic Content Analysis of Litigation EventS (SCALES) - OKN**: Led by Northwestern University, the SCALES open knowledge network is designed to be a public resource to help provide insights based on judicial court records. SCALES is creating tools to decode court records and transform this data into actionable information that aids a variety of uses, including legal scholars, journalists, policymakers, judiciary, and citizens.
- **Scalable Precision medicine Open Knowledge Engine (SPOKE)**: Led by University of California, San Francisco, this Biomedical Open Knowledge Network connects millions of biomedical facts including molecules, pharmacological compounds, organs and diseases, food nutrients, and more. Centered around knowledge representation and reasoning, the team is developing applications using graph theory, advanced visualizations, and real-world clinical evidence to advance drug development and precision medicine.
- **Urban Flooding Open Knowledge Network (UF OKN)**: Led by the University of Cincinnati, the UFOKN is addressing urban flooding impacts to assist decision-makers and urban planners in real-time response and long-term planning.

- **KnowWhereGraph (KWG)**: Led by the University of California, Santa Barbara, KWG provides knowledge graph and geo-enrichment services for environmental intelligence applications. The solution enriches data with pre-integrated custom-tailored knowledge about any locale of interest, thereby reducing the time to find, combine, and reuse data. The initial application areas are focused on decision support related to food systems, supply chains, and humanitarian aid, but can easily be expanded to other application areas as well.
- **Knowledge Network Programming Infrastructure (KNPI)**: Led by the University of Michigan/ MIT, KNPI is building infrastructure for constructing novel OKNs and OKN-powered applications. This solution provides tools to make the creation and maintenance of high-quality datasets and apps more cost-effective and more widely accessible.

The many accomplishments of the five Track A - OKN teams that advanced to Phase 2 are described below, including the integration efforts they have underway to help create elements of a single, nonproprietary, shared knowledge infrastructure. You can find information on these efforts and the other topical tracks supported by the Convergence Accelerator in the **Portfolio**.

SCALES - OKN



The SCALES Open Knowledge Network is an AI-powered platform that provides users access to judicial opinions and insights. It has achieved its project goal of making these transparent to anyone seeking this information.

Project goals and accomplishments

The goal of the SCALES OKN is to bring transparency to the systems and processes of the U.S. courts. Transparency will help ensure that systems and processes are fair, efficient, and accurate. The 94 federal courts have a profound effect on the administration of law in the United States. They decide cases that range from civil rights issues to nationwide product liability cases that affect every single American.

However, the reality is that almost everything we know about how the federal judiciary works comes from the written opinions of judges. These opinions are issued in less than 10 percent of cases — creating a vast dark matter of litigation that academics, policymakers, and the public know little about. To accomplish the goal of transparency, we have built an AI-powered data platform. The platform makes the details of the federal judiciary and insights into how it works available and accessible to everyone.

Through the creation of the SCALES OKN we have developed software and analytics that reveal how the system currently works. This transparency has helped legal scholars, legal aid organizations, and the courts themselves better understand the system and how it could be improved. We developed an analysis on how variable court decisions are. The analysis, included in a report targeting courts and lawyers' associations, shows that the variability of court decisions often has negative effects.

Specifically, the decisions on plaintiff requests to waive the filing fees burden limited-income plaintiffs who cannot pay them. A case in the district courts and the Chicago Council of Lawyers and Appleseed Network have used this report to advocate for a policy change in how these requests are treated. We have also prepared and delivered this report directly to two other federal district courts.

Similarly, we have examined the frequency of motions to seal in intellectual property cases. We also looked at how this frequency varies across districts. The New York University Technology, Law, and Policy Clinic and the Electronic Frontier Foundation included our findings in their brief regarding sealing practices in the courts.

The technology that we have developed to ingest court data now supports a number of other projects. The Civil Rights Clearinghouse is a searchable resource for information and documents relating to civil rights litigation and now uses SCALES software to automatically ingest court records and documents. Similarly, the Full Disclosure Project at the National Association of Criminal Defense Lawyers relies on the SCALES technology stack to automatically download federal court records and populate the project database. The Full Disclosure Project is built to help criminal defense lawyers have a searchable and connected database of police misconduct that aids them in advocating for their clients in court.

Importantly, the goal of the SCALES OKN is to build the connective tissue from court records to other data resources. This will enrich the OKN and advance what users can glean from the data. Towards that end, we have integrated the Federal Judicial Center's open data on judge biographies and metadata for filed cases. We've also created the linkages between the corporate parties involved in cases with the Security and Exchange Commission's EDGAR database. These connections allow users to ask and answer intricate questions about who is involved in a case and how these cases reach a resolution.

Videos, websites, and other resources

The following resources provide additional information about SCALES. It also provides a link for users to register for SCALES to access and explore its information.

- Website: **SCALES OKN**

- **Registration for SCALES OKN access**

- **Science: How to build a more open justice system**
  A policy forum article published in Science by the SCALES team. It describes why systematic analysis of court records is needed. It also describes how it can be accomplished in a manner that benefits the judiciary and the public.

- **ACM Digital Library. From data to information: automating data science to explore the U.S. court system**
  An article that describes the development of the SCALES OKN application prototype and its function.

- **IEEE Abstract. PRESIDE: A Judge Entity Recognition and Disambiguation Model for US District Court Records**
  An article that describes the SCALES OKN models to identify judges by name. Models use free-text. They disambiguate judges' names into a single true entity (i.e., so that all mentions of Judge Jane Doe are connected across all cases).

- **Wiley Online Library. AI Magazine: The Promise of AI in an Open Justice System**
  An article that describes what the SCALES OKN is. It details how AI-driven tools will grow in utility in the legal data domain and help improve the justice system overall.

- **Tear Down This Judicial Paywall - WSJ**
  Opinion from SCALES team members in the *Wall Street Journal.* They advocate for free access to federal court data.

- **Opinion: Texans shouldn't have to pay for court records**
  Opinion from Amy Sanders in the *Houston Chronicle*. Texans should have free access to court records to assess local issues related to social justice.

- **Scales-dash**: *Analysis of Sealing Activity in Patent Cases*
  A public interactive report on the variation frequency of court documents sealing across courts in intellectual property cases.

- **Scales-dash**: *Crosswalking Pacer to the IDB*
  A public interactive report that documents the issues in the official Federal Judicial Center Integrated DataBase when the data column is concerned with issues that affect indigent parties.

- **Civil Rights Litigation Clearinghouse**
  The Civil Rights Litigation Clearinghouse has integrated the SCALES software to automatically download data from PACER into their platform.

- **GitHub - scales-okn/PACER-tools**
  A public repository of our open source software to obtain and parse PACER data. This software has been used directly by multiple federal public defender's offices and researchers.

- **Materials related to SCALES OKN research pieces and blog posts**
  A public repository of all code and data used to produce research publications.

Track integration goals and accomplishments

Our goal is to integrate the work done by KNPS on data-provenance tracking with the Saytrn system developed at SCALES. Satyrn is an information system designed to answer questions about a knowledge network. It does this by mapping information requests from the user into a data analysis and results presentation. Importantly, the system also analytically infers what the proper form of the answer is (whether that be textual or a graph) and delivers the corresponding result. The focus of the work is to provide users with frictionless access to both the information that can be:

- Derived from a data set, and
- Qualified based on the meta-data about its provenance.

The concrete software engineering goal is to integrate the SCALES Satyrn analytics environment with provenance data enabled by the KNPS system. This integration effort would make systems more useful and the resulting software work will comprise two components:

1. Automate Satyrn configuration by exploiting past KNPS provenance recordings, thereby saving human work for each novel data analysis.
2. Provide easy log-driven KNPS provenance publishing hooks in Satyrn so downstream users can reliably review past data analyses.

This software work will be valuable on its own merits, while also furthering higher-level goals:

- Demonstrating implicit work sharing:
  - Among provenance users, and
  - Low-overhead provenance recording by app developers.

The initial work on this integration has focused on defining the structure of the meta-data that will be shared between these two systems and the identification of integration entry points. Our starting point in this work is to use the KNPS metadata as an additional source in the configuration of Satryn. Our longer-term goal is to integrate the configuration layer into the KNPS system itself. The integration will include hooks.

This work is crucial for the broader Convergence Accelerator cohort. The fully realized integration of the Satyrn and KNPS systems will support the effective ingestion, integration, and maintenance of a wide range of data sets. Users, even those without data or analytics skills, can directly access the information that is contained within those data and their analysis. The resulting platform could open the door to a wide range of data sets. Users who would otherwise have to rely on external technical skills or staff could access data sets for more effective utilization.

We believe that this is an opportunity to explore and shape the future of data infrastructure. Data are increasingly utilized in many facets of daily life, and yet the processes for data ingestion, cleaning, reconciliation, and analysis are often bespoke and opaque. As an example, we recall the information challenges of the early stages of the COVID-19 pandemic. Public health officials needed to know the location and status of infections, medical infrastructure, and healthcare workers. While the necessary data existed, mechanisms for aggregating, normalizing, and explaining them did not. We believe this project can help to establish the standards for tools and workflows that would obviate such problems.

SPOKE

SPOKE — the Scalable Precision medicine Open Knowledge Engine — is a biomedical knowledge network. It was created by integrating several specialized databases. It is to be further developed into a multi-domain network of biomedical knowledge and data on a massive scale.

Project goals and accomplishments

Human physiology and pathology are governed by molecular pathways of daunting complexity. These processes are represented in disparate scientific datasets siloed across thousands of public repositories. Siloed datasets make it nearly impossible for researchers to:

- Utilize the entire body of data and factual knowledge, and
- Connect the dots across the domains of specialization in biomedicine.

Yet, Big Data must be converted into Big Knowledge if we are to harness the data revolution. The SPOKE team is working to develop a multi-domain network of both biomedical knowledge and data at a massive scale. These will be validated with real clinical evidence that will enable:

- Investigation of connected biomedical concepts, and
- The emergence of new knowledge

This will be accomplished by the convergence of seemingly disparate knowledge repositories and data sources.

As part of the NSF Convergence Accelerator program, SPOKE integrated several specialized databases. Integrating these databases resulted in a biomedical knowledge network. That network represents 23 million concepts with 50 million connecting relationships. The network is accessible both programmatically (via API) and through a graphical user interface. This allows investigators to reason across a vast, dispersed body of scientifically accepted pathways governing human health. The multidisciplinary project team is collaborating on:

- Network architecture
- Data modeling, and
- Scaled visualizations

The team is also collaborating on applications of the network. These include research questions in the areas of drug discovery and disease diagnosis and management.

The team has demonstrated the utility of the network and the validity of a network-based analysis approach. This approach enables the network to generate hypotheses and provide prognostic information about patients. For example, in March of 2020 SPOKE integrated the newly published SARS-

CoV-2 Interactome into the network to explore some potential pathways of viral activity. They were able to identify pathways and predict promising therapeutic measures. These were later confirmed experimentally in the literature.

The team also launched a company, Mate Bioservices, to commercialize development of a suite of products powered by SPOKE. This suite of products will enable wide dissemination. It will maximize the utilization of this biomedical OKN for the benefit of society as a whole.

By the end of the project period, we will release a web-accessible network visualization tool aimed at citizen scientists. We will produce a report outlining recommendations for mitigating risk associated with ethical, legal, and social implications of network use. And we will deploy the first clinical decision support system (alpha) in the UCSF neurology practice with real patients.

Videos, websites, and other resources

The following resources provide additional information about SPOKE.

- Website: **SPOKE Informational**
- SPOKE **Neighborhood Explorer**
- **Mate Bioservices**
- **A5: Biomedical Open Knowledge Network // Project Video**
- YouTube: **Biomedical researcher**
- YouTube: **Pharmaceutical developer**
- YouTube: **Clinical service provider**

Track integration goals and accomplishments

SPOKE and KnowWhereGraph (KWG) are working on a functional integration plan that will enable users to navigate from one KG to the other when knowledge on that specific area is supported by the partner graph.

While KWG has formal models (i.e., ontologies) for hazards and environmental data layers, including storms, floods, road networks, and health statistics of an area's population, it does not provide mechanistic information on how these could affect individual and population health. Interestingly, this

is exactly what the SPOKE graph contains, providing deep information and connections across diseases, symptoms, genes, drugs, and beyond.

This API-based integration will enable a full service and graph stack for events on the ground and the needs of the affected population. It will drastically improve the matching of experts. At the same time, this project will render both project graphs even more powerful. It will do this by integrating region-specific public health information with detailed data about diseases, their etiology, and recommended treatments. For instance, SPOKE would immediately gain access to making discoveries that involve regional, environmental, and social determinants of health while KWG could add in-depth medical knowledge to its use cases around humanitarian relief and food supply/safety.

Successful API-based integration between SPOKE and KWG will require three steps:

1. Alignment of vocabularies (e.g., taxonomies and ontologies, also called KG schema) used in both graphs.
2. Coreference resolution: Identification of in-common graph nodes and declare their global identifiers (URLs) as equivalent.
3. Establishing a workflow that enables actual queries to their public endpoints, as well as for the usage of these endpoints.

Access to contents across both graphs will provide users with a substantially deeper pool of facts to discover and query both vertically (by adding more data layers) as well as horizontally (by containing more data per layer). In addition, closely integrating the semantic representations (ontologies and design patterns) of the domains within our graphs will enhance interoperability, as well as tie abstract knowledge (such as the Disease Ontology) to concrete data about places in ways that are relevant and immediately impactful for users.

Urban Flooding Open Knowledge Network

The Urban Flooding Open Knowledge Network (UFOKN) is an open and shared infrastructure that enables optimal disaster mitigation and long-term resilience planning.

Project goals and accomplishments

Modern cities are a complex and interconnected system of engineered, natural and social systems. This interconnected system can be conceptualized as a network of networks, or a Multiplex. The multiplex includes the following:

- Power grid
- Transportation network
- Natural surface water and groundwater systems
- Sewerage and drinking water systems, and
- Inland navigation and dams

All are intertwined with the socioeconomic and public health sectors. While the Urban Multiplex connected, the data that have been collected for decades across multiple sectors have remained siloed. That is the main barrier to optimal disaster mitigation and long-term resilience planning. The Urban Flooding Open Knowledge Network (UFOKN) is breaking these data silos. It is an open and shared infrastructure that provides an information backbone for owners, operators, and consumers of various subsystems of the Urban Multiplex during flooding, and for long-term resilience planning.

UFOKN currently holds data on over 140 million critical assets across the continental US including buildings (residential, commercial, industrial, air/water/rail terminals), underground fuel storage tanks, superfunds and the power grid, with more data (e.g. roads and highways, communications infrastructure, wastewater outflow locations, socioeconomic data) being continuously added.

We now have reached the capability to generate real-time flood forecasts at these assets across the continental U.S. These achievements enable critical value adds such as a socioeconomic analysis of regional flood impacts and evacuation routing recommendations. To this end, we have developed a Computable General Equilibrium (CGE) model that assesses the broader economic impacts of flooding on economic activities. These include industry output, employment, and value-added and household income. We generate evacuation routing recommendations using real-time flood forecasts and an agent-based model.

The next phase of the UFOKN project will see a public launch of a series of data products aimed at a wide range of users — from emergency responders to urban planners, utility managers, local and state governments, researchers and the general public.

Videos, websites, and other resources

The following resources provide additional information about UFOKN.

- Website: **Urban Flooding Open Knowledge Network**
- YouTube channel: **UFOKN - YouTube**
- Twitter: **Urban Flooding Open Knowledge Network**
- Article: **Knowledge graphs to support real-time flood impact evaluation - Johnson - 2022 - AI Magazine - Wiley Online Library**

Track integration goals and accomplishments

The way hazards impact communities, infrastructure, and the natural environment depends to a large degree on their connectivity, as well as on their antecedent conditions. For example, the 2018 Camp Fire — the deadliest, most destructive and expensive disaster that year worldwide — was triggered by a small fault in the power grid. But another major contributor, which did not get as much attention, was the prolonged regional drought. The fire destroyed the infrastructure in the region, including buildings, utility networks, and healthcare systems. It caused cascading infrastructural and socio-economic disintegration across urban and rural communities. A large proportion of those displaced from the devastated rural community of Paradise moved to the nearby city of Chico. This increased the population by 20% virtually overnight, overwhelming all available resources.

In order to predict and mitigate such cascading events, we need information systems that integrate and harmonize data from disparate sources. They must also provide actionable information to decision-makers from local and regional to state levels. UFOKN and KWG are collaborating to develop such a system.

UFOKN holds high-resolution data (sub-kilometer, feature-level (e.g., exact location of buildings), while KWG's holdings are at a resolution that is fine-grained from an environmental intelligence perspective, but low for building-level impact modeling in UFOKN. We will develop bridges across geospatial data of varying resolution at a continental US scale. Second, we will develop a common-hazards ontology based on the current KWG version that supports a more fine-grained axiomatization of extreme events and their impacts. Collectively, this work will produce capacity to connect OKNs across various spatial and temporal resolutions and filter relevant connections and knowledge quickly and at scale. We will demonstrate this with a use case based on historic events during Hurricane Harvey in 2017, which we expect will open up further development of ontologies for a broader community.

KnowWhereGraph

The KnowWhereGraph is a cross-domain knowledge graph that supports data-driven analytics and decision-making pertaining to natural disasters and other threats to the environment, industry, and the financial sector.

Project goals and accomplishments

KnowWhereGraph aims to provide area briefings for any place on Earth within seconds. It will answer questions such as "What is here?", "What happened here before?", "Who knows more?", and "How does this region or event compare to other regions or past events?" As a cross-domain, FAIR principles-based knowledge graph for environmental intelligence applications, our current pilots include:

- Disaster relief
- Supply chain management
- Commodity trading
- Financial risk assessment
- Environmental, Social, and Corporate Governance

Our KnowWhereGraph supports data-driven analytics and decision-making by providing (1) a 12 billion facts-strong open knowledge graph that interlinks over 30 cross-domain data layers, (2) a pattern-based suite of expressive ontologies, and (3) a set of geo-enrichment services that enable rapid access to the graph from within the comfort of Geographic Information Systems such as ArcGIS and QGIS.

The team has made significant progress over the past two years in developing the KWG graph using expressive ontologies that connect: multi-source data in terms of disaster, air quality, climate hazards, crop history, soil characteristics, experts and expertise, administrative boundaries, health, transportation infrastructure, and so forth.

Overall, our graph provides 10 different kinds of geographic identifiers and over 20 data layers that provide millions of past and present facts about any of these regions, be it cities, lakes, or agricultural fields. We expect this number to continue growing as more automated graph generation and integration approaches are being developed.

Our team members are also pioneers in developing spatially explicit machine-learning models to provide GeoAI-ready data to empower intelligent decision-making. Combining classical deduction and constraints that checks representation learning enables us to serve a wide range of services to our partners. Services include similarity search, outlier detection, enrichment of existing data, alignment to other graphs, recommendations, link prediction and so on.

Recently, we centered our efforts on supporting project verticals, including the disaster relief subteam,

to assemble quickly needed datasets for rapid disaster response and evacuation after major devastating events, such as hurricanes, have occurred. We are also developing graph solutions for understanding and sustaining food supply-chain resilience.

Each of our pilots comes with bespoke user and query interfaces to accelerate the transition from data to knowledge. We also serve the entire graph openly in the form of a SPARQL query endpoint for developers as well as via our faceted search interface for easy exploration. As a technology-driven project, our goal is to demonstrate how novel geospatial solutions can inform downstream stakeholders from industry, nonprofits, and government agencies.

We are now deploying our graph and services to an increasing set of partners and looking for new opportunities to apply our methods to new use cases.

Videos, websites, and other resources

The following resources provide additional information about the KnowWhereGraph.

- Website: **KnowWhereGraph**

- **KWG - Pilots**
  Pilot programs utilizing KWG

- **KWG - Tools**

- **Know the Graph**
  Link to the graph, faceted search, and query endpoint

- **Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence - Janowicz - 2022 - AI Magazine - Wiley Online Library**
  Janowicz, K., P. Hitzler, W. Li, D. Rehberger, M. Schildhauer, R. Zhu, C. Shimizu, et al."Know, Know Where, KnowWhereGraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence." AI Magazine 43 (2022): 30–39. **https://doi.org/10.1002/aaai.12043**.

Track integration goals and accomplishments

KnowWhereGraph and SPOKE are working on a functional integration that will enable users to navigate from one knowledge graph to the other when knowledge on a specific area is supported by the partner graph. This effort has already been described above under the SPOKE project.

Access to contents across both graphs will provide users with a substantially deeper pool of facts to discover and query both vertically (by adding more data layers) as well as horizontally (by containing more data per layer). In addition, closely integrating the semantic representations (ontologies and design patterns) of the domains within our graphs will enhance interoperability, as well as tie abstract knowledge (such as the Disease Ontology) to concrete data about places in ways that are relevant and immediately impactful for users.

KnowWhereGraph and Urban Flooding OKN are collaborating to develop an information system that integrates and harmonizes data from disparate sources, and that provides actionable information to decision-makers from local and regional to state levels. This joint effort was already described above under the UFOKN project.

Knowledge Network Programming Infrastructure (KNPI)

OKN Infrastructure

Knowledge Networks are a novel and potentially transformative form of data, but building applications on top of them is too difficult, time-consuming, and expensive. We are building a Knowledge Network Programming Infrastructure that makes it far easier to build novel knowledge-powered applications, while also improving the knowledge resources themselves.

Project goals and accomplishments

Open Knowledge Networks are compelling artifacts that enable a range of novel applications, many of which are embodied in the ideas in the Innovation Sprint. Despite OKNs' recent growth, building a novel knowledge network requires substantial engineering effort. Not as many people and organizations benefit from OKNs as should be possible. The core objective of the Knowledge Network Programming Infrastructure project is to make Knowledge Networks easier to build and debug. We do that in several ways:

- Extraction infrastructure software, which makes it easier to populate an OKN with the large set of facts that can be derived from documents
- Provenance infrastructure software, which makes it easier to debug OKNs by creating a queryable record of data operations that is low-cost and works across institutions
- Program synthesis tools that make it easier to induce new OKNs

KNPI extraction software

This software focuses on obtaining OKN facts from natural language documents, in particular scientific documents. Scientific information extraction has been an area of research and engineering activity for at least several decades, but has historically required enormous effort to obtain high-precision and high-recall results. Our extraction software aims to yield high-quality results while dramatically reducing the amount of human effort required. The KNPI extraction suite has four important components.

1. The **VIsualLayout, or VILA**, system is a trained model that breaks documents into visually coherent pieces that are semantically self-contained; for example, article text and image captions will be grouped separately, even if they are geometrically quite close to each other. Previous naive document processors extracted text without concern for visual organization that is obvious to a human observer, making downstream extraction unnecessarily difficult. Using VILA can reduce the error rate of large language model processing by roughly 10%.

2. **Embedding recycling** is a method for accelerating expensive neural model training in the very common setting where a model is repeatedly retrained on an evolving document collection. A large amount of the training procedure exists simply to move parameters into the right rough "neighborhood" for handling the task at hand. This method caches sets of parameters from previous iterations of a trained model, allowing later training attempts to exploit past work without starting

from scratch. On two important tasks for scientific documents — Named Entity Recognition and classification — this method can yield a 55-86% training speedup with only a 0.2% loss in F1 when compared to the non-recycled setting.

3. **Affiliations Linking** is a module for deduplicating organization names that often appear in text with small permutations. For example, a human can tell that the MITt chemical engineering dept, Cambridge, MA, USA, usa refers to the same real-world entity as Massachusetts Institute of Technology (MIT) but not to Massachusetts Chemical Engineering LLC. This problem is endemic in information extraction, and especially problematic when integrating huge numbers of texts into a single resource, as with an OKN. By implementing a pre-ML filtering step that removes obviously improbable candidates, this method is able to obtain linking results that are equal to a conventional Transformer model, at less than half the computational cost.

4. **CascadER Knowledge Graph Link Prediction:** The CascadER system aims to make an ambitious OKN goal — that of adding novel scientific knowledge to an OKN — possible. If we could find the set of potential links in an OKN that are supported using published evidence, it would amount to creating a structured representation of scientific knowledge. Methods exist to do this, both purely from text and (more productively) from multimodal inputs such as text plus figures.

   Unfortunately, state-of-the-art approaches take up to 11 days of processing to evaluate a single novel OKN edge (because they test every possible answer using the entire text as evidence). The CascadER system is structured to evaluate evidence in a particular order, from "coarse and cheap" to "fine and costly". By better using computational resources, it can obtain substantially better quality results than competing methods.

Knowledge Network Provenance System

The Knowledge Network Provenance System (KNPS) software system enables easier data production for OKNs. Provenance systems attempt to capture a record of data operations for later use. If we had a record of all OKN-relevant data processing, we could answer a range of questions that would greatly accelerate development, including:

- What training set was used to generate a particular model?
- Is a dataset suitable for use with medical products?
- Was a particular visualization created by people I trust?

Unfortunately, provenance systems are not widespread enough to capture this information in most circumstances today. Traditional provenance systems — sometimes seen in financial applications, or certain relational databases, or large-scale ML platforms — are characterized by (a) substantial software

modifications and (b) detailed and precise operations of each data record. Point (a) means that collecting provenance information is extremely expensive. Since modern OKN and other data production pipelines often involve a range of teams, datasets, tools, and institutions, even substantial investment in provenance software can be insufficient for capturing the end-to-end information necessary to answer the questions above.

We note that the cost advantages of social OKN dataset construction might allow us to address this problem. OKNs can yield datasets that are much larger and cheaper than traditionally administered databases. They do this by combining machine learning and social mechanisms to make it feasible to ingest large sets of data from a wide variety of sources. The resulting dataset may not be as "perfect" as a traditional database but is often much larger with quite good quality.

Our KNPS provenance system follows this approach, by:

1) Admitting low-quality instrumentation information from desktop clients, cloud service logs, and other scanners
2) Using machine learning and social OKN mechanisms to upgrade this cheaply collected provenance data when appropriate, and to infer missing data when possible.

The current platform has a range of scanners, a graphical front-end for examining the collected data, and a set of mechanisms for inferring and upgrading the low-quality inputs.

CORD-19 is a large OKN containing scientific information about the coronavirus, extracted from more than 1M papers. It was produced using the extraction software described above.

Videos, websites, and other resources

The following resources provide additional information about the KNPS.

- The CORD-19 dataset: **GitHub - allenai/cord19: Get started with CORD-19**
- The KNPS system: **GitHub - mikecafarella/KNP**
- The suite of extraction tools have been used to create a public-facing repository of scientific paper extractions: **Semantic Scholar.**

Track integration goals and accomplishments

Our primary in-track integration work is with the SCALES team. This work is described above in the SCALES section. The primary work from the KNPS perspective lies in creating provenance recordings that can be repurposed for future configuration and deployment. Technically, there exists an interesting opportunity similar to that of large-scale text "foundation models" like GPT-3 or GitHub Copilot that can capture many tasks in a single trained model.

We aim to collect a large number of provenance recordings, first for SCALES, and then for other tools as well. The resulting provenance OKN can be used to construct a model that describes the overall distribution of software configurations, much in the same way that large-scale text models can capture the overall distribution of useful natural language sentences. At later use time, an administrator should simply specify a tiny number of distinctive configuration parameters, and then ask the model to fill in the remaining information.

We have started to collect provenance information, and have ancillary code ready for constructing a trained embedding-style model from OKN information. Once data collection is complete, it should be fairly straightforward to test how accurately we can autoconfigure Satyrn with this statistical record of past usage.

National Science Foundation
Directorate for Technology, Innovation and Partnerships

BETA.NSF.GOV/TIP/LATEST