OPEN KNOWLEDGE NETWORK ROADMAP:

POWERING THE NEXT DATA REVOLUTION

SEPTEMBER 2022

AUTHORS

OKN Innovation Sprint Organizing Committee

Chaitan Baru
University of California, San Diego

Lara Campbell
National Science Foundation

Wo Chang
National Institute of Standards and Technology

Tess DeBlanc-Knowles
Office of Science and Technology Policy /
National Science Foundation

Jemin George
National Science Foundation /
DEVCOM Army Research Laboratory

Martin Halbert
National Science Foundation

Co-Authors

Kat Albrecht
Georgia State University

Luis Amaral
Northwestern University

Nariman Ammar
University of Tennessee

Todd Bacastow
Maxar Technologies

Sergio Baranzini
University of California,
San Francisco

Matt Bishop
Open City Labs

Michael Cafarella
Massachusetts Institute
of Technology

Silviu Cucerzan
Microsoft Corporation

Ying Ding
University of Texas at Austin

Brian Handspicker
Open City Labs

Oktie Hassanzadeh
IBM Research

Pascal Hitzler
Kansas State University

Florence Hudson
Columbia University

Sharat Israni
University of California,
San Francisco

Angela Rizk-Jackson
University of California,
San Francisco

Esther Jackson
Columbia University

Eric Jahn
Alexandria Consulting

Krzystof Janowicz
University of California,
Santa Barbara

Bandana Kar
Oak Ridge National Laboratory

Sam Klein
Massachusetts Institute
of Technology

Matthew Lange
International Center for Food
Ontology Operability Data and
Semantics

Ora Lassila
Amazon

Chengkai Li
University of Texas at Arlington

Ryan McGranaghan
Atmospheric and Space Technology
Research Associates

Murat Omay
U.S. Department of Transportation

Adam Pah
Northwestern University

Co-Authors

Louiqa Raschid
University of Maryland

Glenn Ricart
US Ignite

Emanuel Sallinger
Vienna University of Technology /
Oxford University

Greg Seaton
SierraLogic LLC

Cogan Shimizu
Kansas State University

Amanda Stathopoulos
Northwestern University

Paul Wormeli
Wormeli Consulting LLC

Lilit Yeghiazarian
University of Cincinnati

Ellie Young
Common Action

Technical Writers/Editors

Pamela Livingston
Writing Assistance Inc.

Douglas Maughan
National Science Foundation

Shelby Smith
National Science Foundation

ACKNOWLEDGEMENT

This document attempts to capture the insights of a large and diverse set of stakeholders who shared their time and insights over many months.

The authors are grateful for the enthusiasm and compelling expertise of the participants in the Open Knolwedge Network Innovation Sprint, who from February to June of 2022 attended six workshops and numerous smaller meetings. Many thanks are due to the federal agencies, companies, universities, non-profits, and other organizations who allowed their staff to join this effort. The authors particularly appreciate the enabling work of the NSF Track A: Open Knowledge Network grantees whose work has shown the power and promise of an OKN as well as the need to build a larger community.

The authors also gratefully acknowledge the many prior efforts that are referenced in this document and those that are not. We hope that participants from those prior efforts, participants from the OKN Innovation Sprint, everyone who reads this report, and other interested parties will continue to help advance the effort to create an Open Knowledge Network.

36.7702
30.7317
59.4454

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Open access to shared information is essential for the development and evolution of artificial intelligence (AI) and AI-powered solutions needed to address the complex challenges facing the nation and the world. The Open Knowledge Network (OKN), an interconnected network of knowledge graphs, would provide an essential public-data infrastructure for enabling an AI-driven future. It would facilitate the integration of diverse data needed to develop solutions to drive continued strong economic growth, expand opportunities, and address complex problems from climate change to social equity. The OKN Roadmap describes the key characteristics of the OKN and essential considerations in taking the effort forward in an effective and sustainable manner.

Data, information, and knowledge

Harnessing the vast amounts of data generated in every sphere of life and transforming them into useful, actionable information and knowledge is crucial to the efficient functioning of a modern society. Data and information should be easy to find, access, and reuse. Knowledge structures that enable integration of vast amounts of diverse data in service of a very broad range of uses across society represent the national infrastructure for a prosperous future. *Knowledge graphs* are an important type of knowledge structure to enable such integration. They consist of *nodes* and *edges* — where *nodes* represent real-world entities (e.g., a city, a neighborhood, a court case, a gene, a chemical compound), and edges represent different types of relationships among nodes. The Open Knowledge Network is envisioned as an open, interconnected network of knowledge graphs that serves as public, accessible infrastructure. This infrastructure will enable development of a variety of solutions for a broad spectrum of societal use cases using open, public data, as well as data requiring controlled access.

In February 2022, the National Science Foundation (NSF) in partnership with the White House Office of Science and Technology Policy, launched an *OKN Innovation Sprint*. The Sprint harnessed the collective insights of roughly 150 experts from government, industry, academia, and nonprofit organizations to help build a roadmap for a Prototype OKN (*Proto-OKN*), from specific use cases and end-user perspectives. The main findings from this Sprint are summarized in this report. The context for this activity is provided by the original vision for an OKN described in the NSF Harnessing the Data Revolution Big Idea and by projects in Track A: Open Knowledge Network of the **NSF Convergence Accelerator**.

Discussions during the Innovation Sprint helped clarify that the creation of an OKN is fundamentally a sociotechnical effort that must consider human, social and organizational factors, as well as a technical effort. Deep engagement is necessary among domain knowledge experts and a host of other stakeholders including data owners, decision-makers, various end-user communities, tool builders, and knowledge representation experts. User-centered design approaches, stakeholder involvement and alignment, and customer engagement are essential to ensuring an impactful and sustainable outcome. The Innovation Sprint focused on the needs and requirements of end-users and various stakeholders. Sprint participants formed 17 use-case groups. Thirteen groups focused on "verticals" (i.e., specific application areas), and four focused on cross-cutting, "horizontal" themes. **Section 6** of the report introduces these use cases. Appendix A describes their activities and outcomes in detail.

Report outline

The OKN Roadmap report first introduces the OKN concept (**Section 1**), and prior initiatives related to this effort (**Section 2**). The essential characteristics of the OKN include an open system architecture, a dynamic system reflecting real-world information updates, the ability to link diverse information and deduce linkages among related information elements, and a simultaneous focus on the use case-driven "vertical" aspects as well as the technology-driven "horizontal" aspects. Moreover, critical aspects of the system include a governance structure to allow OKN stakeholders to direct and oversee policies, processes to ensure ethical use, methods to track data provenance, flexibility to scale up to support additional data and users, interoperability with new data sources and systems, sustainability to meet current and future end-user needs, enforcement of data access rights, and procedures and metrics for data validation (**Section 3**).

The activities needed to create the OKN include gathering requirements from various use cases, establishing the necessary information structures such as ontologies and schemas, working with existing repositories to link them to the OKN, fostering development of information structures to facilitate interconnecting data across domains, developing working prototypes of the system for specific use cases, facilitating engagement by subject matter experts, developing a variety of user-friendly interfaces for a broad range of users, and ensuring that it is possible to incorporate public as well as sensitive, access-controlled data. Data and Information flow within the OKN could be considered in several distinct phases which include *design* of appropriate information structures for proper data representation, data *ingestion* into the system, data *enrichment* to extract information from data, data *storage*, data use/*consumption* by various applications defined by different use cases, and on-going maintenance of the whole enterprise (**Section 4**).

In moving forward, the efforts to construct an OKN must be guided by the sociotechnical nature of the undertaking. Strong stakeholder engagement essential for the success of this effort can be built through practices like the participatory design approach that engages stakeholders as well as end-users and developing educational and training materials suited to the broad range of stakeholders. Stakeholder engagements can be further strengthened by enabling open contributions for a range of sources to promote transparency and employing a community-centric approach to ensure an ethical and responsible system. Long-term success of the effort would require building a committed community of stakeholders to secure sustainability, ensuring that new data and end-users are able to connect to and use the OKN, and building the connective fabric of OKN to prevent future fragmentation of the system **Section 5**).

The Sprint participants recognized that the construction and deployment of OKN would be a community-based effort to develop a national-scale data infrastructure with the costs distributed across many stakeholders. The schedule for a Proto-OKN would take into account the possibility of leveraging the available technology and experience base, while also paying attention to the significant new needs and requirements uncovered during the Sprint process.

Looking ahead

The OKN is an essential element of a national AI infrastructure. It would provide users across all sectors — including government agencies, private organizations, academia, and the public — access to integrated information for a variety of uses, including tackling societal problems, driving evidence-based policies, and developing novel AI capabilities.

This is the opportune time to embark upon development of the Proto-OKN by embracing state-of-the-art technologies, leveraging related efforts, and building upon the many partnerships that have developed via the NSF Convergence Accelerator and over the course of the OKN Innovation Sprint.

OPEN KNOWLEDGE NETWORK ROADMAP:

POWERING THE NEXT DATA REVOLUTION

The Open Knowledge Network (OKN) Roadmap paves the way for the rapid deployment of an Open Knowledge Network — an essential infrastructure for enabling an artificial intelligence (AI)-driven future. The open infrastructure would enable sharing of open, public data, and potentially data requiring controlled access. It would provide the knowledge infrastructure necessary for integrating the diverse data needed to continue strong economic growth, expand opportunities, and address complex problems from climate change to misinformation, disruptions from pandemics, and advancing economic equity and diversity. Access to rich, structured data is essential for evolution and use of AI and AI-based methods and solutions to the complex challenges facing society today. The OKN Roadmap describes the key characteristics of the OKN and essential considerations in taking the effort forward in an effective and sustainable manner.

In 2017, the National Science Foundation (NSF), in its **Harnessing the Data Revolution Big Idea**, recognized the need for an open networked structure for linking disparate, heterogeneous information from diverse sources to yield novel data-driven insights and solutions. As a first step, in Fall 2019, NSF funded 21 OKN projects as part of the **Convergence Accelerator Track A** program. In Fall 2020, five of these projects entered a 2-year Phase 2 effort. In February 2022, the NSF and the **Office of Science and Technology Policy (OSTP)** engaged a larger community of subject matter experts, end-users, and stakeholders from government, industry, academia, nonprofits and other communities in a 4-month Open Knowledge Network Innovation Sprint. This activity culminated in a Proto-OKN Roadmap workshop in June 2022, leading to this OKN Roadmap report.

This OKN Roadmap report guides readers through the various considerations in creating and deploying an OKN. It provides background information demonstrating the compelling need for an OKN. It specifies needs and requirements to serve various stakeholder interests, ranging from climate change to the criminal justice system, developed through design exercises focused on a diverse set of use cases. The report is comprised of the following eight sections:

- **Section 1** introduces the vision of the OKN, including its features, functions, and benefits, and related work in this area.
- **Section 2** describes the initiation of the OKN activity and the OKN Innovation Sprint process.
- **Section 3** describes the key takeaways related to the characteristics of the OKN.
- **Section 4** describes the range of issues to be considered in creating an OKN.
- **Section 5** describes considerations for taking this effort forward in an effective and sustainable way.
- **Section 6** provides an overview of 17 use cases developed during the Innovation Sprint.

- Section 7 describes a possible timeline for implementing a Proto-OKN.
- Section 8 provides a conclusion based on the findings of the previous seven sections.
- Appendix A provides a detailed account on all 17 use cases from the OKN Innovation Sprint.
- Appendix B provides an overview of the five OKN Phase 2 projects supported by the NSF Convergence Accelerator's **Track A Program**.

1. Envisioning the OKN

The vision for the OKN is the creation of a common infrastructure that is driven by specific use cases but takes the form of a shared, general platform. The OKN is envisioned to transform our ability to unlock actionable insights from data, by semantically linking information about related entities.

The Internet – which began as an attempt to link files and then evolved into a digital infrastructure that serves as the backbone for modern life – and the related idea of *Semantic Web* (Berners-Lee et al., 2001), can be seen as an inspiration and model for OKN development. The OKN is envisioned as an ethical, trustworthy network of *interconnected knowledge graphs* (Chaudhari et al., 2022). It would provide a trusted platform for users accessing information as well as for users who are information providers by establishing and following ethical standards for data exchange and analysis.

Knowledge graphs — founded on the principle of applying a graph-based abstraction to data — have emerged as a compelling concept for integrating information and extracting value from multiple diverse sources of data at large scale (Hogan et al., 2022). Some of the largest knowledge graphs in existence are powering consumer applications including web search, e-commerce, advertising placement, and question-answering (**Noy et al., 2019**). These same technologies can be used to create an open platform, with open as well as access-controlled data, to develop impactful new applications for evidence-based policymaking, game-changing research, and many other key areas of societal impact.

The benefits include the ability to provide answers to questions that might otherwise require inordinate effort in assembling, integrating, and analyzing datasets curated by different organizations. For example, questions such as those listed below could be answered quickly by the OKN.

- "Have there been unusual clusters of earthquakes in the United States in the past six months?"
- "What is the best combination of chemotherapeutic drugs for a 56-year-old female with stage 3 brain cancer?"

Providing a platform for shared access and cross-sector information-sharing

Envisioned as an inclusive, open, community-driven infrastructure accessible to all, the OKN would not only provide a trusted platform to empower a host of new applications, but it would also open new vistas in AI and data science research, including in fairness, bias, diversity, equity and inclusion. Access to OKN infrastructure and content would allow researchers and practitioners to develop more robust and efficient approaches to answering questions, more expressive frameworks to capture knowledge, and more natural interfaces to access that knowledge (**NSCAI Final Report, 2021**). Any organization regardless of size or sector could benefit via a multi-sector, community-based effort that would help share the burden of development via open-source software and open, shared standards.

Considering factors and approaches for OKN design

Stand up of an OKN would leverage prior work in **Semantic Web** and **Linked Open Data**, and make full use of the experience and the platform provided by the highly successful **Wikidata** effort, as well as other more recent data commons efforts.

Many of the core technologies that underlay the OKN are well-established (Chaudhari et al. 2022; Hitzler 2021; Hogan, et al. 2022; Gutierrez, 2021). However, successful creation of the OKN is much more a sociotechnical challenge (Baxter and Sommerville, 2011) than merely a technical exercise. The design and implementation of an OKN includes human, social and organizational factors, as well as technical factors. Creating the OKN requires deep engagement among domain knowledge experts and a variety of other stakeholders. These include data owners, decision-makers, various end-user communities, tool builders, and data visualization experts.

Launching an Innovation Sprint

Reacting to the various drivers for the OKN, the NSF, in partnership with OSTP, launched an OKN Innovation Sprint to bring together a diverse set of stakeholders to help design a roadmap for a prototype of an OKN (Proto-OKN). Running from February through June 2022, the Innovation Sprint harnessed the collective insights of experts from over 24 academic institutions, 15 Federal agencies, 3 non-profit organizations, and 20 private-sector companies.

Experience from the NSF Convergence Accelerator had made clear that user-centered design approaches, stakeholder involvement and alignment, and customer engagement were essential to ensuring design of an ethical, impactful and, ultimately, sustainable OKN. Thus, the Sprint activities focused on approaching the vision for an OKN through end users and uses of an OKN. The Innovation Sprint concluded that building the OKN would bring significant fiscal and temporal savings as effort and expense for data identification and data use continue to rise. The OKN infrastructure would bolster data-driven decision making, reduce the likelihood of using inaccurate and/or incomplete data, support coherent management and access to data resources, reduce expense for development of standards including ensuring ethical compliance, and enhance service provision for a broad range of use cases.

This report summarizes the main findings of the Innovation Sprint.

2. Origins of the OKN concept

The vision of the OKN has been developing robustly and with broad support over the past five years. The initial vision was put forward in 2017 as part of the NSF Harnessing the Data Revolution Big Idea. The idea was further developed via community meetings, culminating in a workshop organized by the Federal Interagency Big Data Working Group (NITRD OKN workshop). The imperative for an OKN to support the AI research environment was also noted in the final report of the National Security Commission on Artificial Intelligence.

In March 2019, the National Science Foundation issued a Dear Colleague Letter (DCL) announcing the Open Knowledge Network as one of the inaugural Tracks of the NSF Convergence Accelerator. The NSF objective was to engage multidisciplinary, multi-institutional teams to help identify development paths for an OKN, with a particular focus on exploiting publicly available U.S. government and similar public datasets.

In September 2019, the NSF Convergence Accelerator selected 21 multidisciplinary projects for a one-year Phase 1 effort as part of its inaugural OKN Track (Track A). Phase 1 efforts included developing the team's idea into a proof of concept, identifying new team members and partners, and participating in the innovation curriculum. At the end of Phase 1, teams participated in a formal proposal and pitch, which was used to select teams for Phase 2. In September 2020, 5 of these 21 projects were selected for a two-year Phase 2 effort. Phase 2 efforts focused on high-impact deliverables and sustainability. Teams transformed their prototype into solutions and developed sustainability plans to continue impact beyond NSF support. For more details on the Convergence Accelerator OKN Phase 2 projects, please refer to Appendix B.

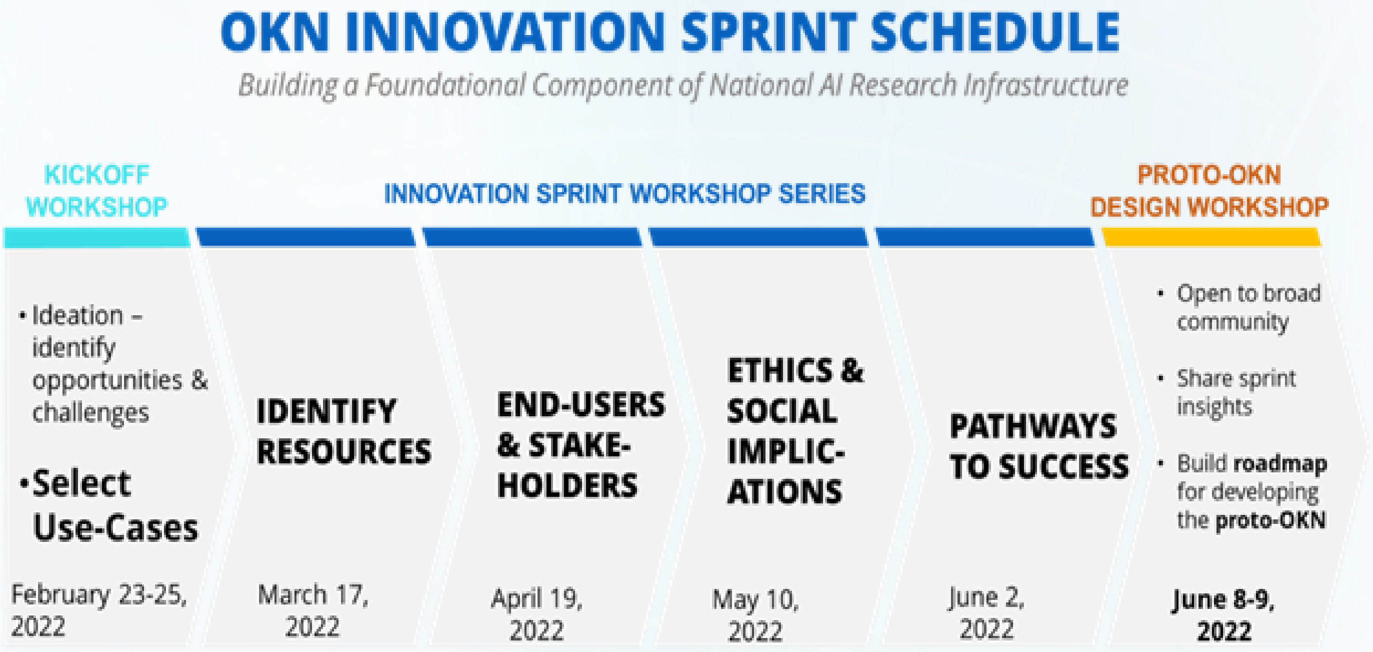
The OKN Track A projects leveraged and benefited from prior work in knowledge graphs and related technologies, including Linked Open Data and the Semantic Web. However, they also uncovered numerous new technical challenges as well as sociotechnical considerations. In November 2021, the NSF issued a Dear Colleague Letter focusing on the research issues uncovered by OKN efforts (CISE OKN DCL).

These prior efforts provided the context for the OKN Innovation Sprint. This report represents the collective findings and recommendations of the Sprint contributors. It serves as a guide for future, iterative Proto-OKN development.

OKN Innovation Sprint

The goal of the OKN Innovation Sprint was to assess the opportunity space for creating an OKN and the possible challenges in its formation. The Sprint kickoff meeting on February 23-25, 2022, attracted over 150 attendees representing all sectors — government, academia, industry, and nonprofits. Community members signed up for the 4-month sprint exercise and organized themselves into 17 different groups representing a wide range of application areas and themes. **Figure 1** displays Sprint activities from the kick-off meeting to the Sprint's conclusion.

Each group identified essential characteristics and components of an OKN and then created specific use cases to help guide its further development. Each group consisted of two self-volunteered group leads, several members, and an active Slack channel. Groups met weekly from March through June 2022, and monthly with the Innovation Sprint Organizing Committee. The use cases resulting from their weekly design efforts are detailed in Appendix A of this report.



OKN INNOVATION SPRINT SCHEDULE

Building a Foundational Component of National AI Research Infrastructure

KICKOFF WORKSHOP

- Ideation – identify opportunities & challenges
- Select Use-Cases

February 23-25, 2022

INNOVATION SPRINT WORKSHOP SERIES

IDENTIFY RESOURCES

March 17, 2022

END-USERS & STAKE-HOLDERS

April 19, 2022

ETHICS & SOCIAL IMPLIC-ATIONS

May 10, 2022

PATHWAYS TO SUCCESS

June 2, 2022

PROTO-OKN DESIGN WORKSHOP

- Open to broad community
- Share sprint insights
- Build **roadmap** for developing the **proto-OKN**

June 8-9, 2022

Figure 1: Sprint schedule steps, by date

The Sprint activity concluded with a capstone Proto-OKN Design Workshop on June 8-9, 2022. Each group presented results from their Sprint activities and outputs from all groups were pooled together to serve as the basis for the Proto-OKN design. The OKN Innovation Sprint exercise revealed the tremendous excitement and commitment by the community for the OKN vision and approach and the broad scope of opportunities that it affords.

3. Characteristics of an OKN

The Innovation Sprint revealed the need to engage a broad range of stakeholders and end-users in order to obtain a holistic view of the characteristics of the OKN. Stakeholders ranged from data curators and data programmers to a variety of end users. The Sprint exercise helped capture their needs and requirements as part of the design process.

A user-centered design approach is necessary to engage stakeholders early in the development process. User-centered design shows stakeholders the benefits that they and their clients or customers will receive from the proposed end-product. Designing an OKN that accommodates the needs and preferences of users is essential for its successful adoption. In a project intended to provide easy access to a wide range of users with differing needs, failing to accommodate users' needs will limit the project's success, making it more difficult to sustain over a long period of time.

The Innovation Sprint defined a number of essential characteristics of an OKN, including:

- An open architecture to allow inputs from a variety of sources.
- A dynamic system reflecting real-world information updates and changes as they occur.
- Ability to link disparate information by traversing links across the network and to deduce linkages among entities.
- Interlinking of "horizontal" and "vertical" elements of the network. "Horizontal" elements provide a common technological infrastructure and sociotechnical frameworks, agnostic to the knowledge domains. "Vertical" elements focus on preparation and ingestion of content from specific information domains.

The Innovation Sprint also identified seven elements critical to the establishment and sustainment of an OKN:

- Governance: a set of guiding principles to enable OKN stakeholders to direct and oversee policies, design, development, operations, and management of the system.
- Ethics: structures and processes based on ethical principles that provide consistency, transparency, and accountability.
- Provenance: methods that are able to track any and all changes to information over time, including original sourcing of an asset and subsequent processes that are employed.
- Scalability/Interoperability: a flexible and extensible data infrastructure that is *elastic* and enables the exchange, use, and transfer of OKN resources.
- Sustainability: an extendable ecosystem that enables the OKN to meet current and future needs with backward and forward compatibility.
- Access Rights: a flexible system to enable access to OKN resources from open public access to predefined privileged access and analysis.
- Data Validation: procedures, processes, and metrics for ensuring that ingested data are accurate, complete, and meet specified criteria, including distinguishing trustworthy data from untrustworthy data or misinformation.

An OKN created by integrating a variety of datasets held by diverse entities would:

- Facilitate answering dynamic, interconnected questions for a variety of users, e.g., government employees, companies, private citizens, nonprofits.
- Increase flexibility for ongoing data integration plus future use and data re-use.
- Empower AI and machine learning tools by creating pre-integrated, AI-ready data.
- Enable data-driven insights to tackle the "unknown unknowns" and improve decision-making capabilities for agencies and organizations of all sizes.
- Reduce the expense of data cleaning and "data wrangling."
- Reduce entry hurdles by basing the OKN framework on open, well-established Web standards, allowing use of off-the-shelf, open-source tools.
- Enhance data interoperability by moving from software and platform dependency to semantically interconnected data.

4. Creating an OKN

The OKN is envisioned as an open, domain-agnostic, community-based effort for establishing a national-scale data infrastructure, with the development and maintenance costs being distributed across many stakeholders. The management and maintenance of the OKN is also expected to be a distributed, editorialized process. Established collective intelligence efforts like *Wikipedia* and *Wikidata* provide a model and valuable experience base as a starting point for OKN.

Another important experience base to build upon is the design thinking and user-centered design approaches that have been employed in the NSF Convergence Accelerator Program. These design approaches have proved essential in identifying user needs and requirements, and to develop prototypes and the so-called Minimum Viable Products (MVPs). This approach has not only helped clarify implementation issues and priorities but has also served to engage and empower a diverse set of stakeholders in the process.

The Innovation Sprint identified the activities needed to create a prototype OKN, which include:

- Gathering requirements for use-case capabilities specific to each domain, aggregating capabilities as general requirements by their corresponding horizontal requirements, understanding which of these capabilities are immediately available versus which would require research advances, and identifying the types of linkage required among data from different domains, e.g., health and environment, natural resources, or judicial records.
- Establishing quality schemas in the form of ontologies while taking existing ontologies into account and identifying any gaps, compiling inventories of relevant data resources, services and frameworks across different domains, adopting common or shared representation.
- Encouraging existing repositories to provide easy communication among the prototype-OKN, **NIEM ontologies,** and extensions of the **schema.org** framework wherever possible. Identifying and addressing any barriers to access would help the OKN create more robust data access to various domain repositories, including private and sensitive data.
- Fostering interconnection of information across domains, with particular emphasis on those that may currently be largely disconnected and/or difficult to integrate.
- Developing highly effective prototypes for querying and accessing data and, where applicable, performing reasoning tasks with the data.
- Enabling involvement of domain/subject experts without substantial technical skills, e.g., to verify/validate/curate the knowledge base.
- Prototyping various user-friendly interfaces with different data access modalities that enable expert, non-expert, as well as non-technical users to access and use the data, information, and OKN services.
- Developing metrics to objectively measure the use and the impact of OKN use by different stakeholders.
- Ensuring that it is possible to source data from a wide variety of resources including unstructured, semi-structured, and structured sources and, importantly, data with varying levels of quality and fidelity.
- Developing an approach so that the open system design of the OKN would also be able to support

the ability to incorporate private and access-controlled data using established governance principles and procedures.

The flow of data and information in the OKN system could be considered in several distinct phases: *Design, Ingestion, Enrichment, Storage, Consumption*, and *Maintenance*.

- *Design* would consist of a description of target use cases with requirements and assessment of relevant data sources. An ontology and schema of the knowledge graph for the given use case would then be developed based on these descriptions and assessments. The schema should be developed, documented and made available alongside the graph data to support future updates and evolution of the graph and its use cases.
- *Ingestion* would include setting up data pipelines and workflows to construct/create the graph according to the design. This includes registering and establishing access to external data sources, mapping "legacy data" to the knowledge graph structures, and incorporating data-quality checking measures for each data source.
- *Enrichment* would consist of a set of methods designed to increase quality and applicability of the graph using techniques such as entity disambiguation, aligning/mapping schemas and data to existing ontologies and graphs, incorporating measures to help increase data quality, and making data integration decisions explicit, reusable, and reversible. Enrichment could also include the use of knowledge management methods for completing data-predicting links, discovering links to other OKNs, and assessing data quality. The assessments may include identifying inaccurate and contradictory data, creating derived data, representing different contexts for the data, and obtaining different semantic views to serve different stakeholder/user communities.
- *Storage* would include use of state-of-the-art technologies for managing the ontologies and the knowledge graphs themselves using state-of-the-art storage techniques including graph databases and triple stores using cloud-based implementations. Storage and access via the cloud would help establish OKN as a widely shared infrastructure.
- *Consumption* would typically be via a variety of user interfaces as well as software interfaces, namely application programming interfaces (APIs), for tasks including knowledge graph access, ontology management, provenance/lineage management, and other re-use and administration tasks. User interfaces/APIs would be required for a wide range of functions, including OKN curation, data administration, data exploration and querying of information. Software interfaces (APIs) would be based on open standards (e.g., the W3C "Semantic Web stack" and other related specifications) to allow third parties to build applications using the OKN.
- *Maintenance* of the schema and graph would be secured on an ongoing long-term basis. Suitable versioning (e.g., through distinct releases) would be used to minimize disruptions to applications depending on the graph.

The approach to building the OKN system should be iterative and evolving, to facilitate flexibility and adaptation of the system to changing needs. A set of measurable metrics should be developed to help gauge functionality, performance, and progress. These would be reinforced through continual user testing.

There should be a centralized, lightweight governance board, with a hierarchical structure that enables distributed data governance at the dataset level. The governance structure would include a "data working group" to help interface with federal agencies and other data holders, and to help these stakeholders integrate their data. To ensure an ethical system, the system governance should adhere and adapt to current frameworks and guidelines, including supporting open standards and **FAIR** data (Findable, Accessible, Interoperable, and Reusable) while regularly evaluating the system for potential harms.

5. Considerations moving forward

The OKN represents a best-in-class opportunity to provide **FAIR access** to open data, to enable AI and ML tools and ecosystems, and to leverage data and information needed to address societal challenges and innovation opportunities. The OKN Roadmap provides guidance for how a cross-government effort for Proto-OKN development should proceed. The prototype should strive to provide a public data infrastructure for the use of government agencies and other stakeholders. The prototype would demonstrate the ability to integrate and use public data to develop novel solutions to some of today's most complex challenges facing Americans. It would also serve to enhance public access to data resources created with American taxpayer dollars.

As already mentioned, the OKN would leverage the technological advances and experience base in several areas including the semantic web; ontologies and ontology engineering; the Wikidata ecosystem; Linked Open Data, and others. It would leverage available technologies and existing standards including those from **W3C**, **IEEE**, **OASIS**, and **NIEM**. At the same time, this activity will also help define new standards, processes, and methods in technical areas such as standardizing workflows/processes for data ingestion and for data enrichment, as well as in the sociotechnical aspects of creating a national infrastructure like the OKN.

Strong stakeholder engagement is essential for the success of OKN, and it would be built around the following key considerations:

- *Participatory Design*. OKN would be developed on human-centered design principles that engage stakeholders and end users. Importantly, this would include development of education and training materials that address a broad audience — from technical staff to end users and senior management, to the general public.
- *Open contributions*. To avoid the risk of a closed, proprietary system that may benefit only a few, the system would be open to all. Transparency is an essential attribute of a trustworthy system. The data, metadata, and the processes employed for data ingestion, enrichment and consumption would be open, accessible, and transparent.
- *Ethics*. Ensuring an ethical, responsible system requires an intentional and community-centric approach and careful design, along with ongoing monitoring. Many factors will contribute to the success of this objective, including working directly with communities and ensuring inclusivity in the data, users, and communities engaged.
- *Sustainability*. An open, inclusive system requires a committed community of stakeholders to ensure

sustainability of the purpose, data and network. Starting by interconnecting use-cases of clear value to agencies and other organizations will help build that community, as will connecting to existing open science efforts, data library initiatives, and shared research computing structures. Creation of or inclusion in a coordinating entity such as the proposed National AI Research Resource could be valuable for developing the long-term support needed for operation and maintenance of OKN as an infrastructure.

- *Extensibility*. The OKN should be architected using existing standards described above, but with an eye toward the future. New data, use cases, and partnerships must be able to connect to and use the OKN to provide more value and impact.
- *Connectivity*. Maintaining a focus on the sources of data, end-users of OKN tools, and other stakeholders is essential. The human-centered design inherent in developing the use-cases and building the connective fabric of OKN should help prevent fragmentation.

6. Use cases

The OKN Innovation Sprint activities approached the OKN design process through the lens of a variety of use cases, each focusing on a specific societal need. Participants formed 17 use case groups. Thirteen of those groups focused on specific use cases, or “vertical” applications, of the OKN, and four focused on cross-cutting, “horizontal” themes. The 13 “vertical” use cases fall in the broad categories of:

- Equity, Social Care, and Justice Issues
- Climate Change, Disaster, Energy Systems
- Health Communications and Information Accuracy
- Innovation and Research Ecosystems, and
- Macro-level issues, including:
    - Supply Chain Information
    - Data-driven Decision Support, and
    - Financial Risk Analysis.

A more detailed discussion of each use case is provided in Appendix A. Each use case in the following narrative bears the name of the group assembled to create it and by the factors necessitating its creation.

Equity, social care, and justice issues

The following three OKN use cases focus on the concept of providing community care to improve community health and well-being for all community members. Community care would be delivered through social services such as decarceration service planning and family reunification; addressing and preventing homelessness; and increasing transparency in the justice system.

Group B: Integrated Justice Platform

This use case group explored the data infrastructure needed to collect, aggregate, and harmonize data across domains of the justice system in order to improve the way it works, reveal patterns and trends across data systems, and evaluate the influence of bias and other extralegal factors on institutions, communities, and individuals.

Group H: Homelessness

This group explored requirements for creating a knowledge graph that would provide near real-time data tracking of homelessness, available housing, and shelter and social services usage. It would also track related funding and programs to help city and community leaders house unhoused citizens and prevent future homelessness.

Group M: Decarceration Service Planning System

This group focused on development of OKN-based social services that would enable persons released from incarceration to successfully return to their communities and families. These services would help reduce recidivism and improve the well-being of formerly incarcerated individuals and their families.

Climate Change, Disaster, Energy Systems

The four use cases in this group are broadly related to climate change and its impact on food systems, various aspects of local communities and community-level decision making, energy systems, and on natural disasters which, due to climate change and other factors, have routinely become compounded events.

Group E: Food and Climate

This group focused on developing a food and climate use case. They determined that applying existing climate and food-system ontologies to public-agency award data would improve investment outcomes and enhance understanding among investment decision-makers. Ontologies could then provide access to more comprehensive information. Decision-makers accessing this information would be better able to make informed decisions about where food investments would provide the greatest benefit. They would also be able to more easily identify investment gaps.

Group G: Climate Change

This group focused on a multi-faceted decision-support system that uses a coherent set of climate data to help local to global community residents and decision-makers make informed decisions. Such decisions have the potential for significant impact on community health and infrastructure, criminal-justice system equity, and environmental protection. Poorly made decisions could harm individuals, communities, and the environment. Decisions made with the support of the proposed OKN-based decision-support system could help. The OKN would mitigate a broad range of issues and enhance resilience by revealing interrelationship among social, economic, and environmental factors.

Group L: Energy

This group explored development of an OKN-based decision-support tool that would enable the Department of Energy, public utilities, and local communities/stakeholders to:

- Assess risks to the grid from extreme events
- Determine community energy resilience, energy justice, and burden, and
- Identify opportunities for renewables to reduce energy burdens while achieving the objectives of clean energy.

Group P: Compounding Disaster Events

This group explored the use of an OKN to aggregate and connect diverse, multi-scale, distributed data. They looked at how a knowledge graph of this kind could deliver prognostic and real-time information to decision-makers before and during disaster events. Such events tend to co-occur, resulting in complex impact scenarios that require coherent responses from a wide range of stakeholders. In evolving, uncertain contexts, poor decisions could be made without knowledge-graph support. And climate-change impacts could be made worse.

Health Communications and Information Accuracy

The use cases for this topic explored how bona fide public health information could be made available to counter the significant amounts of misinformation on this topic in social media, especially focusing on vulnerable populations in this context.

Group F: Health-Related Information

This group explored the development of a data infrastructure for health-related information — as well as misinformation — with the goal of mitigating health misinformation and improving public health. The OKN developed to meet these goals would promote healthy behaviors, especially among vulnerable populations, such as older adults, low-income families, and minority communities.

Innovation and Research Ecosystems

Use cases for this topic explore how to enable data-driven, insightful decision-making support for stakeholders of various research and innovation programs. Specifically, they explored how to enable the system to support researchers seeking to identify current research gaps and opportunities.

Group C: Defense Innovation Programs

This group investigated construction of an OKN-based repository of all innovation programs currently funded by the U.S. Army and other Department of Defense agencies. The proposed repository would help users identify linkages among various programs, eliminate duplication of effort, initiate new collaborations, and accelerate the transition of technology at scale to end users.

Group R: Gaps in Research

This group investigated how development of an OKN that aggregates micro-level data on research activity could help to identify overlooked research approaches; find gaps in research as targets for future work; and facilitate a more equitable and robust allocation of funding and better target human resource training.

Supply Chain, Decision Support, Financial Risk Analysis

These use cases address macro-scale issues in supply chain management, decision support, and financial risk assessment. They promote the use of data-driven insights, decisions, and predictions for use by public and private agents to tackle the issues listed below.

Group A: Strategic Supply Chain

This group considered development of an OKN for several uses including real-time visibility of supply chain flows (e.g., for water, masks, and medical supply); identification and prediction of bottlenecks and shortages; crisis management and disaster recovery; sustainability strategies; and industry health. Supply-chain resilience is essential for each of the above.

Group J: Decision Support for Government Leaders

This group investigated development of an OKN-driven decision-support system at federal, state, and local levels to provide citizens and decision-makers data-driven insights and predictions related to a broad range of decisions. These pertain to the following:

- Land-use
- Transportation options
- Clean air and water regulation, and
- Emergency and disaster response.

Group Q: Public Industry Risk Analysis and Alert System

This group focused on creating a business and finance OKN curated from publicly available sources. The knowledge graph would provide indicators and indices pertaining to opportunities and risk factors. Publicly available sources could include:

- SEC filings
- Analyst reports, news reports, and Web text

Cross-cutting ("Horizontal") Issues

The use cases for this topic explored common technological infrastructures and sociotechnical frameworks that could be used to create the OKN.

Group D: NIEM Ontology

This group explored the use of the National Information Exchange Model (NIEM) to build an ontology that can serve as a foundation for other knowledge-graph design efforts and sample knowledge graphs. The ontology would help achieve faster cross-domain interoperability by leveraging the underlying ontology. This would improve integration and interconnectedness among knowledge graphs/networks across a broad variety of scientific domains and governmental entities.

Group I: Electronic Consent Services

This group investigated the possibility of developing an electronic consent service as a common mechanism for managing privacy preferences. The consent service would cover consent assertions for protecting personal and/or sensitive information in an OKN-based data-sharing system.

Group K: Collaborative Knowledge Graph for Researchers

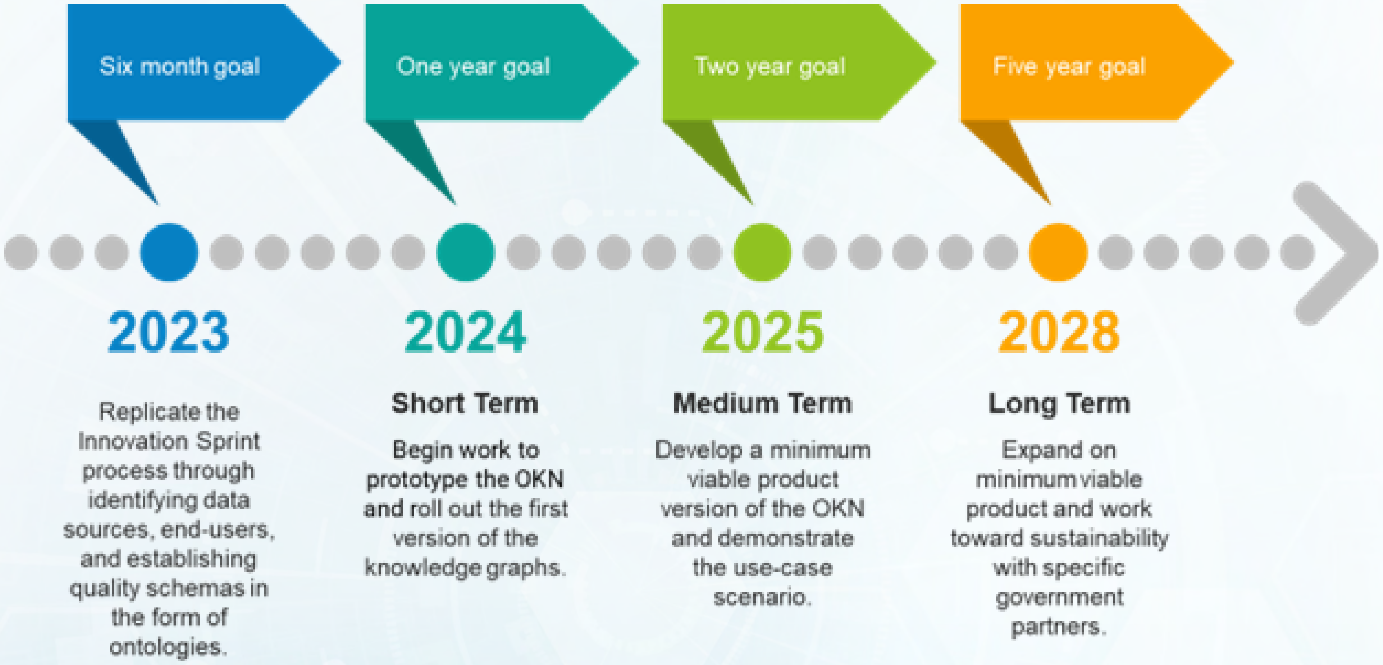
This group explored a distributed software environment for the +OKN where participants could establish their own “nodes” or leverage pre-existing infrastructure to publish, access, and consume knowledge graphs with privacy and ethical safeguards in place.

Group N: Learning Resources

This group focused on the needs and requirements for an OKN Training and Education Gateway to provide learning and outreach materials to the entire spectrum of stakeholders from agency administrators and elected officials, to technical/software architects, data contributors, professionals, educators, researchers, students, and members of the public.

7. Schedule and Effort

The OKN effort should leverage existing work in related areas and the current technological base. The use of user-centered design methods and iterative implementation methodology will be important in order to match evolving needs and requirements of various use cases. Prior efforts like the NSF Convergence Accelerator OKN Phase 2 and Wikidata should be recognized and leveraged, while also paying attention to the significant new needs and requirements uncovered during the Innovation Sprint. The schedule should take these factors into account while developing the goals, objectives, and deliverables for the short, medium, and longer term (i.e., over 6 months, 2 years, and 5 years). **Figure 2**, below, displays the goals for years 2023 through 2028.



Replicate the Innovation Sprint process through identifying data sources, end-users, and establishing quality schemas in the form of ontologies.

Begin work to prototype the OKN and roll out the first version of the knowledge graphs.

Develop a minimum viable product version of the OKN and demonstrate the use-case scenario.

Expand on minimum viable product and work toward sustainability with specific government partners.

Figure 2: Four phases of OKN development over 5 years

8. Conclusion

Creating the OKN as a network of interconnected knowledge graphs will power the next data revolution by enabling the use of vast troves of data and information to:

- Develop and share real-world knowledge
- Accelerate collaborative innovation, and
- Address current and future societal challenges

The ever-increasing generation of digital data and the need to integrate these data for a broad range of applications with societal impact requires a platform like the OKN. Knowledge representation at the scale envisioned by OKN is essential to the success of future AI-based methods. As a national-scale infrastructure the OKN can benefit a broad range of constituents including government agencies, private organizations, non-profits, academics, and others. It can drive a broad range of novel applications aimed at tackling societal challenges, including use of open data for evidence-based policies, and also help in developing novel AI capabilities.

Publicly supported OKN-related activities have made significant progress over the past few years beginning with Track A in the **NSF Convergence Accelerator** (see Appendix B). The OKN Innovation Sprint has generated considerable community interest and enthusiasm because the OKN holds tremendous potential to impact transformative change not only within individual sectors and domains, but for the nation as a whole. This is the opportune time to embark upon the creation of OKN, beginning with a Proto-OKN development activity.

References

Baxter, G. and Sommerville, I., Socio-technical systems: From design methods to systems engineering, *Interacting with Computers*, Vol. 23, No. 1, January 2011, pp. 4–17.

Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web, *Scientific American*, Vol. 284, No. 5, May 2001, pp. 34-43.

Chaudri, V., Baru, C., Chittar, N., Dong, X., Genesereth, M., Hendler, J., Kalyanpur, A., Lenat, D., Sequeda, J., Vrandečić, D., and Wang, K., Knowledge Graphs: Introduction, History and Perspectives. *AI Magazine*, Vol. 43, No. 1, March 2022, pp. 17-29.

Guiterrez, C. and Sequeda, J., Knowledge Graphs. *Communications of the ACM*, **Vol. 64, No. 3, March 2021, pp. 96-104**.

Hogan, A., Guiterrez, C., Cochcz, M., de Melo, G., Kirranc, S., Pollcrcs, A., Navigli, R., Ngonga Ngomo, A-C., Rashid, S. M., Schmclzciscn, L., Staab, S., Blomqvist, E., d/Amato, C., Labra Gayo, J. E., Ncumaicr, S., Rula, A., Scqucda, J. and Zimmermann, A., **Knowledge Graphs**. SpringerLink, 2022.

Hitzler, P., A Review of the Semantic Web Field, *Communications of the ACM*, Vol. 64, No. 2, February 2021, pp. 76-82.

National Science Foundation (NSF), "**Harnessing the Data Revolution at NSF**", 2020.

National Security Commission of Artificial Intelligence (NSCAI). **The Final Report**, 2021.

Noy, N., Gao, Y., Jain, A., Narayan, A., Patterson and A. Taylor, J., Industry-scale knowledge graphs: lessons and challenges, *Communications of the ACM*, Vol. 62 No. 8, July 2019, pp. 36-43.

36.7702
30.7317
59.4454

National Science Foundation
Directorate for Technology, Innovation and Partnerships

BETA.NSF.GOV/TIP/LATEST