OPEN KNOWLEDGE NETWORK

SUMMARY OF THE BIG DATA IWG WORKSHOP
OCTOBER 4–5, 2017

Product of the

BIG DATA INTERAGENCY WORKING GROUP

SUBCOMMITTEE ON NETWORKING & INFORMATION TECHNOLOGY
RESEARCH & DEVELOPMENT

COMMITTEE ON SCIENCE & TECHNOLOGY ENTERPRISE

of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

NOVEMBER 2018

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure that science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the Networking and Information Technology Research and Development Program

The Networking and Information Technology Research and Development (NITRD) Program is the Nation’s primary source of federally funded work on pioneering information technologies (IT) in computing, networking, and software. The multiagency NITRD Program, guided by the NITRD Subcommittee of the NSTC Committee on Science and Technology Enterprise, seeks to provide the research and development (R&D) foundations for assuring continued U.S. technological leadership and meeting the needs of the Nation for advanced IT. The National Coordination Office (NCO) supports the NITRD Subcommittee and the Interagency Working Groups (IWGs) that report to it. More information is available at https://www.nitrd.gov/about/.

About the Big Data Interagency Working Group

The NITRD Big Data IWG focuses on R&D to improve the management and analysis of large-scale data to develop the ability to extract knowledge and insight from large, diverse, and disparate data sources, including mechanisms for data capture, curation, management, and access. The Big Data IWG works to identify current big data R&D activities across the Federal Government and to offer opportunities for coordination among agencies, academia, and the private sector. More information about the Big Data IWG is available at https://www.nitrd.gov/nitrdgroups/.

Acknowledgments

This workshop report was developed through contributions of the WSRD workshop committee represented by members from government, industry, and academia; NITRD Federal agency representatives and members of the WSRD IWG; and the staff of the NITRD NCO. Sincere thanks and appreciation go to all who contributed.

Copyright Information

Background

Technology companies develop proprietary knowledge networks as key business technologies today. However, because these networks are proprietary and expensive to construct, government, academia, small businesses, and nonprofits do not have access to them. In contrast, an "open" knowledge network (OKN) would be available to all stakeholders, including the researchers who will help push this technology further. An OKN requires a nonproprietary, public–private development effort that spans the entire data science community and results in an open, shared infrastructure. This infrastructure has the potential to drive innovation across science, engineering, and finance and to achieve economic growth comparable to the impact of the Internet in the early 1990s.

Just as the Internet began as an attempt to link files and then became a major infrastructure, the OKN is intended to link data about related entities. For example, many government agencies have been investing in efforts to create specialized knowledge networks in domain-specific areas such as genomics, astronomy, physics, or the geosciences. However, fusing these islands of knowledge currently requires enormous effort. Like the Internet, an OKN would provide such an infrastructure, building upon and significantly enhancing the capabilities of existing data resources.

We are witnessing the first wave of this technology in consumer "conversational" knowledge services (e.g., intelligent virtual assistants). However, these services are limited in scope and closed to contributors beyond their corporate firewalls. As a result, they can only answer limited questions in their respective business areas. Such "walled gardens," like internet portal efforts in the 1990s, typically do not fare as well as open efforts like the World Wide Web.

An OKN would enable machine learning systems to enrich data with extensive information about the underlying objects and natural language systems to link words and sentences to meaningful descriptions and add context and world knowledge to robotic devices. With OKN access, academic researchers could develop more robust and efficient approaches to answering questions, more expressive frameworks to capture knowledge, and more natural interfaces to access that knowledge. All companies, regardless of size or sector, would be able to take advantage of the OKN. Major technology companies could potentially share the cost burden of development by following models such as open source software ecosystems or efforts to create shared search engine standards.

Why Now?

Commercial Efforts

In the commercial sector, knowledge graphs or ontologies—representations of interlinked descriptions of "entities", i.e., real-world objects, events, situations, or concepts—have demonstrated and are powering new capabilities for search and integrated services. The technology exists for expanding to thousands of new topic areas and classes of questions. Commercial successes present the opportunity for an effort to drive the next generation of this technology to support research in academia, industry, and government, and to create new services and synergies that use both open and proprietary data.

International Efforts

International efforts related to OKN are underway. The European Open Science Cloud initiative and its implementation strategy GO FAIR aim to catalyze major investments with a focus on scientific knowledge. In China, major corporations are collaborating with universities and investing heavily in knowledge graphs. Canada's data cyberinfrastructure initiative includes efforts to build linked open datasets and knowledge graphs. Collaborations with partners in Canada, the European Union, the United Kingdom, and Japan could strengthen international OKN efforts.

Key Takeaways

The Networking and Information Technology Research and Development Program (NITRD) Big Data Interagency Working Group (BD IWG) hosted a workshop October 4–5, 2017, in Washington, DC, on the *Open Knowledge Network*. The purpose of the workshop was to bring together Federal, industry, and academic stakeholders to discuss issues related to developing a national-scale semantic information infrastructure. Workshop breakout sessions discussed OKN issues for a horizontal, technological domain and for four specific vertical domains: biomedicine, geosciences, finance, and smart manufacturing. A detailed workshop agenda, list of attendees, presentations, and breakout group reports are available at https://www.nitrd.gov/nitrdgroups/index.php?title=Open_Knowledge_Network.

Workshop participants acknowledged that an effective OKN could drive the next wave of knowledge-powered artificial intelligence (AI) and transform domains ranging from scientific research to commercial applications by enabling services such as recommendation systems, translation systems, social media services, and intelligent search agents.

Participants identified the following key characteristics of an OKN:

- **Dynamic in nature** to reflect real world information updates and changes as they occur.
- **"Open"** to accommodate input from a variety of sources.
- **Able to link disparate information** by traversing links across the network and by deducing linkages among entities.
- **Comprised of both "horizontal" and "vertical" elements:**
  - ***Horizontal activities*** enable a common technological infrastructure, regardless of the knowledge domains; these capabilities should include:
    - Query services
    - Integration services
    - User-friendly ("natural") interfaces
  - ***Vertical activities*** prepare the content for the OKN. Tasks may be domain-specific but will include such things as mapping ontologies to the OKN and extracting information from structured and/or unstructured sources for inclusion in the OKN. Key vertical activities include:
    - Collecting representative sets of queries
    - Compiling inventories
    - Designing prototypes
    - Enabling the interconnection of data from disparate sources
    - Developing metrics

Workshop Sessions

Overview of Agency Presentations

Presentations by Federal agencies focused on OKN-related projects they have undertaken or currently are pursuing, and the importance of an OKN to their missions. The National Science Foundation (NSF) envisions OKN as a component of its Harnessing the Data Revolution "Big Idea" that with industry input could foster an entirely new class of application that leverages data, context, and inferences from data. For several years, the Defense Advanced Research Projects Agency (DARPA) has been making OKN-relevant resources available, such as catalogs, testbeds, and software, as well as "Data-to-Wisdom" program investigations. The National Aeronautics and Space Administration (NASA) has been exploring methods to create, use, and maintain knowledge graphs, including how they enable the search for information and data at large scales. The OKN fits with the National Institutes of Health (NIH) strategic vision for managing data under the FAIR (findable, accessible, interoperable, and reusable) framework, using policies, standards, repositories, and incentives to support open data and open science. The National Institute of Standards and Technology (NIST) is interested in the standardization aspects of an OKN, the development of the necessary technical ontologies, and the economic opportunities it would enable.

Overviews and Summaries of Breakout Sessions

The Horizontal Domain

Overview: This session discussed the basic technical approaches and technologies that should span all content domains of an OKN effort. In practice, many existing knowledge graphs are implemented as "entity stores."[1] The speakers laid out the key technical elements of an OKN, including cataloging of entities using unique identifiers; matching engines to link text with entities; representation of facts as entity–relationship–entity "triples" augmented by *provenance* and *timestamp*; and normalization engines to handle ambiguity. They highlighted the point that data, regardless of quantity, has limited utility if it is not well integrated. For example, data.gov has ~178,000 data sets from many government organizations and NIH-supported DataMed has ~1.5 million medical data sets.[2] Yet each of these data sets stands in isolation; it is not easy to integrate them or to navigate across collections. OKN could provide the means for treating all distributed data sources as one—similar to Web search engines that enable access to the entire distributed Web as if it were a single site.

Attendees noted a variety of questions and applications that a fully operational OKN could potentially support. For example, Molecular Tumor Boards[3] could use an OKN to address cancer treatment questions such as, *"Treatment so far has not worked; what pending drugs or recent research papers are relevant to this patient's glioblastoma, given her specific genetic mutations, markers, and family history?"*

Other examples of questions across a breadth of domains include:

- Which Hodgkin Lymphoma treatments for my mother are covered under her insurance?
- What do the cells in capillary systems of liver tumors unresponsive to sorafenib have in common?
- What are good things to do with kids in Pittsburgh?
- Which Washington-based think tanks have worked on projects involving South American trade?
- What is the address of the building I am in? Where do I go for a taxi?

Examples of applications include:

- Application (App) Development: To develop a useful app for domain X, which needs entities in domain Y (e.g., a great cancer app should be able to furnish users bus routes to treatment centers).
- Machine Learning: To allow secondary and tertiary features and aggregates to be used in machine learning algorithms.
- Robotics: To incorporate common sense reasoning for robots to understand, not simply sense, their environments.

[1] In computer science, an entity is anything about which information can be stored in a database; for example, a person, concept, physical object, or event.

[2] See https://www.data.gov/ and https://datamed.org/.

[3] Tumor boards are meetings where physicians caring for patients with cancer and other providers meet to discuss specific patients and advise one another on the best treatment plans.

- Analysis: To support the scientist or analyst who needs to reconcile the data across disparate data sets to use them for analysis.

Summary: The horizontal (technology) breakout session identified the following as critical factors in the development of an OKN prototype:

- Adoption of a "triple" representation format along with *provenance* and *timestamp* as the basic representation scheme.
- A query server (httpd equivalent[4]) that incorporates reasoning capabilities.
- An integration or fusion server (equivalent to a search engine) with reasoning capability.
- A Web-browser-like natural language interface that is pointed at data.

In addition, authentication was called out as a "horizontal issue," which if solved for one use case or domain could be used across all OKN domains.

The Biomedical Domain

Overview: The Biomedical session discussed the complexity of biomedical information and the need for an OKN to deal with this complexity. An academic knowledge network for determining new drug treatment alternatives was presented. In this effort, drug treatment alternatives for specific conditions such as nicotine dependence could be based on pathways between 11 node types (including genetics, compound types, and side effects) and 2.25 million relationship arcs. This network continues to grow with the addition of epigenomic and environmental factors and larger experimental compounds databases important to drug discovery.

Summary: The key actions identified in the Biomedical breakout session were to:

- Bring together siloed ontologies, data resources, frameworks, and projects via a shared representation.
- Set objective measures for OKN effectiveness.
- Develop meaningful but tractable use cases.

Ontologies are a "low-hanging fruit" in this domain. An example is NIH's Semantic MedLine,[5] with 25 million articles, where even partial input, say 50%, of its data would be a good basis for building an OKN Prototype.

One idea that was expressed is to start with three well-established ontologies and three major use cases. For example, the GO (gene),[6] UniProt (protein),[7] and BioPortal (cell-related)[8] ontology projects (funded in part or in total by NIH) were proposed. A possible use case query would be offering alternate treatment pathways to clinicians.

[4] httpd, or http daemon, is a software program that runs in the background of a web server to wait for and automatically answer incoming server requests.

[5] https://skr3.nlm.nih.gov/SemMedDB/

[6] http://www.geneontology.org/

[7] https://www.uniprot.org/help/gene_ontology

[8] https://bioportal.bioontology.org/ontologies/CL

The Geoscience Domain

Overview: The Geoscience session identified several resources relevant to the OKN effort. These resources included community ontologies and standards with broad coverage, e.g., NASA's "top-level" Semantic Web for Earth and Environmental Terminology (SWEET) ontology[9] (standard sensor model language), or other sensor ontologies for time and space.[10] Infrastructure components—such as geoscience data centers with semantic application programming interfaces (APIs), and the Earth Science Information Partner (ESIP) Community Ontology Repository[11]—were mentioned as examples of semantic data in geoscience. The integration of natural and human models, key to research in sustainability and management of natural resources, was noted as an area in critical need of OKN-like capabilities. Participants suggested an initial focus of an OKN could be the NSF EarthCube initiative.[12]

Summary: Participants in the Geoscience breakout session discussed the relevance of geoscience information in OKN not just for scientists (who currently invest enormous effort to integrate data) but for a much larger and broader range of users, such as policymakers, businesses, resource managers (e.g., for water resources), first responders, and the public. Specific challenges for OKN that are posed by geosciences include the grounding of data in space and time; the need to store large datasets (e.g., 3D grids); the pervasive uncertainty in data values that stems from measuring and modeling; and the lack of clear spatial and temporal boundaries for many objects of interest, e.g., a storm, a bay, or a fault.

Several short-term steps were proposed:

- Understand the nature and requirements of geoscience queries from both the general public (e.g., "what was the trend in air quality last month in Bethesda, MD?") and scientists (e.g., "find fisheries data for the last 50 years in the upper Hudson river").
- Encourage existing geoscience data repositories to share their data to prepopulate ("seed") an OKN, perhaps through extensions of existing community schema activities.
- Connect geosciences data with other kinds of data in OKN, such as health and financial data.

Proposals for midterm steps for geosciences included natural language interfaces for geospatial and temporal aspects of OKN; tools for assessing data quality and uncertainty; connecting data explicitly with entities (e.g., scientists who collect the data, data centers that serve it, and software to process it); handling real-time data in OKN; and addressing what-if scenario queries.

A potential driving scenario in geosciences could be flooding events, because (1) they combine a wide range of scientific model data with Federal and local data about a region, (2) the results are of interest to first responders as well as to commercial and volunteer organizations and the general public, (3) there is urgency in data publication and dissemination and on quick integration and reuse that can be facilitated by the semantic infrastructure provided by OKN, and (4) there are important societal impacts and current interest, given recent severe flooding events.

The Finance Domain

Overview: The Finance session noted that classic finance problems include determining whether a given entity is independent of or interdependent with other entities, and working out the relationships

[9] https://sweet.jpl.nasa.gov/

[10] https://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7908.pdf

[11] http://esipfed.org/; several Federal agencies support ESIP, including the Environmental Protection Agency, NASA, the National Oceanic and Atmospheric Administration, and the U.S. Geological Survey.

[12] https://www.nsf.gov/geo/earthcube/

among entities (e.g., from traditional financial statements and market transactions, one can identify "contracts" as a primary entity). There also may be several types of relationships among entities at several levels (e.g., *genetic, phenotypic,* and *population*, to borrow terminology from the biosciences.) With such a structure, the execution of a contract can be viewed as an automaton or structured as data. Although rather late to adopt semantic infrastructure, the finance community initiated a challenge in 2016 for entity identification with finance data. A case study used the residential mortgage-backed securities issue at the heart of the U.S. 2008 financial crisis. Text analytics applied to public prospectuses were used to construct the financial supply chain for these securities. This unique dataset was then used to track the impact of toxic financial entities that issue subprime mortgages on the downstream performance of the securities.

Summary: Participants in the Finance breakout session recognized that in comparison to the Biomedical or Geoscience domains, the U.S. financial communities are relatively late in adopting semantic infrastructures that are the precursors to the OKN. Incentives for an open network need to be carefully studied and aligned with current business models to clearly demonstrate benefits to some of the user community. A possible use case might be to utilize OKN to make small businesses more competitive. For example, many companies provide "Know Your Entity" services, and one could use OKN to develop a similar capability for small businesses so that users could obtain business intelligence relevant to their sectors. Example OKN queries were suggested, such as, "Is company Z too big to fail?" and "What assets belong to company Y?"

A near-term goal may be to input relevant data collections, e.g., the Securities and Exchange Commission EDGAR database[13] or the Federal Reserve Board FRED database.[14] There is a need for a small set of simple ontologies, e.g., related to basic finance concepts, financial events, and financial contract types. The finance community initiated an annual Financial Entity Identification and Information Integration (FEIII) Data Challenge that in 2018 focused on the creation of a simple OKN to capture knowledge of companies in the S&P 500 index in two different industry sectors.[15] The challenge was to rank all the competitors of a seed financial entity. In addition, the group also identified several mid- to long-term tasks, e.g., to enhance the North American Industry Classification System classification of financial entities, and to identify complex relationships among and associated with financial events.

The Smart Manufacturing Domain

Overview: The Smart Manufacturing session discussed the need for a system-level classification of the main terms used in manufacturing—for instance, in describing products or patents—as a first step towards integration of manufacturing data. Five use cases for OKN were described: manufacturing capabilities, products, patents, uses of robots/sensors, and interoperability for smart manufacturing.

Summary: The Smart Manufacturing breakout group proposed that a demonstration OKN focus on sensors, including metadata about types of sensors and the types of data captured or emitted by them. Sensor data can be used to describe, diagnose, predict, and prescribe, and they may be incorporated in catalogs not routinely searched by web search engines. The Clean Energy Smart Manufacturing Innovation Institute could provide resources, including catalogs and real-world data, and support for pilot projects.[16]

[13] Electronic Data Gathering, Analysis, and Retrieval system, https://www.sec.gov/edgar.shtml

[14] Federal Reserve Economic Data, https://fred.stlouisfed.org/

[15] https://ir.nist.gov/feiii/2018-challenge.html

[16] The Clean Energy Smart Manufacturing Innovation Institute (CESMII) is the ninth manufacturing institute awarded under the Department of Energy; https://www.cesmii.org/manufacturing-usa/.

The initial focus of a sensors OKN pilot project would be identifying existing OKN infrastructure for manufacturing, including selected catalogs and big datasets. This step would be followed up by developing some sample applications to demonstrate payoff from a sensors OKN and then extending the applications to other manufacturing domains.

Overall, attendees observed that there could be significant benefits if the OKN was populated simultaneously across all these application areas. One could jump across areas and perform entity resolution and search for information crossing domain boundaries (in the style of Wikipedia). This would be an exciting new capability, because in the past these areas have been hard to traverse.

Moving forward: Action Items and Research Considerations

Build an OKN Community Across Government, Academic, and Industry Stakeholders

Discussions identified several possible paths toward building a broad-based OKN research community:

- **Hold several types of workshops:**
  - A planning workshop to identify initial goals of a broad-based OKN research community and define prototypes or challenge problems in a variety of vertical domains.
  - A series of workshops, one for each vertical domain, including a final, consolidation workshop.
- **Conduct outreach to improve communication between communities**, including encouraging communities to inform funding agencies about their activities to promote the OKN:
  - Develop a manifesto listing the purpose, contours or outline, and stewardship of the OKN.
  - Publish articles in domain-specific journals and premier business publications.
  - Engage with industry consortia to find industry partners that can collaborate on specific OKN efforts, e.g., to develop applications.
- **Develop and implement curricula in semantic technologies at various levels**, e.g., undergraduate, graduate, and training.

Research Considerations

The Horizontal Domain

Participants widely considered three major capabilities to be essential to the “horizontal” capabilities of the OKN:

1. **Query services** requiring the design and prototyping of “query servers” to enable easier access to OKN data and other content, but most importantly, for performing reasoning tasks with the data.
2. **Integration services** requiring design and prototyping of “integration servers” able to access and “integrate” data and information from distributed resources via a single point of access.
3. **User-friendly (“natural”) interfaces** requiring the design and prototyping of natural language interfaces to enable expert, non-expert, and nontechnical users to access data, information, and services available in the OKN. OKN is clearly a *data*-based representation; just as it would have been difficult to communicate the value of the Web without the first web browser, there is a need for a natural language interface for data that can demonstrate OKN’s value to potential users.

The Vertical Domains

Participants broadly discussed a set of activities for the biomedicine, geosciences, finance, and smart manufacturing vertical domains; all activities would require engagement with the larger community associated with each vertical:

- **Collect representative sets of queries in each domain**, classifying those queries by the corresponding horizontal capabilities required; understanding which capabilities would require research advances; and identifying the types of linking required among data from different domains, e.g., health and environment, natural resources, or smart manufacturing.
- **Compile inventories in each domain** of existing ontologies, data resources, services, and frameworks, with the goal of using a common or shared representation to enable inclusion in the OKN; this activity would facilitate broader understanding of existing resources in each domain.
- **Foster development of prototypes**, seeding prototype OKN efforts in the communities as well as encouraging existing repositories to provide access to their data to seed an OKN, perhaps through extensions of the *schema.org* framework.
- **Foster interconnection of information** among domains, especially those that may currently be largely disconnected and difficult to integrate.
- **Develop metrics** that can define objective measures of OKN use by different stakeholders, adopt various technologies and vocabularies, and effectively support growth with minimal effort.

Conclusion

Artificial intelligence, machine learning, natural language technologies, and robotics are all driving innovation in information systems. Developing the knowledge bases, graphs, and networks that lie at the heart of these systems is expensive and tends to be domain specific, and the largest currently are focused on consumer products (e.g., for web search, advertising placement, and question answering). An open and broad community effort to develop a national-scale data infrastructure—an Open Knowledge Network—would distribute the development expense, be accessible to a broad group of stakeholders, and be domain-agnostic. This infrastructure has the potential to drive innovation across medicine, science, engineering, and finance, and achieve a new round of explosive scientific and economic growth not seen since the adoption of the Internet.