International Coordination for Science Data Infrastructure: Proceedings of a Workshop in Brief

DETAILS

8 pages | 8.5 x 11 | null
ISBN null | DOI 10.17226/25015

CONTRIBUTORS

Board on Research Data and Information; Policy and Global Affairs; National Academies of Sciences, Engineering, and Medicine

Proceedings of a Workshop

IN BRIEF

FEBRUARY 2018

International Coordination for Science Data Infrastructure

Proceedings of a Workshop—in Brief

Advances in science and technology have led to the creation of large amounts of data—data that could be harnessed to improve productivity, cure disease, and address many other critical issues. Consensus in the scientific community is growing that the transition to truly data-driven and open science is best achieved by the establishment of a globally interoperable research infrastructure.

A number of projects are looking to establish this infrastructure and exploit data to its fullest potential. Several projects in the United States, Europe, and China have made significant strides toward this effort. The goal of these projects is to make research data *findable*, *accessible*, *interoperable*, and *reusable*, or FAIR (see Box 1). The expected impact and benefits of FAIR data are substantial. To realize these benefits, there is a need to examine critical success factors for implementation, including training of a new generation of data experts to provide the necessary capacity.

On **November 1, 2017**, the Board on Research Data and Information (BRDI) of the National Academies of Sciences, Engineering, and Medicine organized a symposium to explore these issues. Invited experts from China, Europe, and the United States were asked to:

- Review proposed science data infrastructure projects around the globe;
- Highlight, compare, and contrast the plans and capabilities of these projects; and
- Discuss the critical success factors for implementation and the role of international cooperation for scientific data management.

BOX 1

The FAIR Data Principles set out requirements for the sharing of scientific data

Findable: Easy to find by both humans and computer systems and based on mandatory description of the metadata that allows the discovery of interesting datasets.

Accessible: Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content.

Interoperable: Ready to be combined with other datasets by humans as well as computer systems.

Reusable: Ready to be used for future research and to be processed further using computational methods.

Source: Barend Mons, presentation, November 1, 2017, Washington, DC. Credit given to the Dutch Techcentre for Life Sciences.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

SETTING THE STAGE

BRDI chair **Alexa McCray**, Harvard Medical School, welcomed participants from federal agencies, academia, the private sector, nongovernmental and international organizations, professional societies, and the philanthropic community. **George Strawn**, BRDI Director, likened today's information infrastructure issues to issues around the creation of a network infrastructure 25 to 30 years ago. Just as technical and social structures came together to create what we now know as the Internet, technical and social issues must be addressed for a robust information infrastructure. Symposium chairman **James Hendler**, Rensselaer Polytechnic Institute, noted the move away from asking *why* data coordination should occur to focusing on *how* to do it, what efforts are underway, and what pitfalls to avoid. For this reason, the symposium was organized to begin with presentations about ongoing efforts, followed by newer initiatives and research perspectives to build international structures for information sharing.

GO FAIR: TOWARD THE INTERNET OR FAIR DATA AND SERVICES

Bernard Mons, Leiden University Medical Center and chair of the European Open Science Cloud (EOSC), focused on the political aspects of open science. Development of the Internet solved the problem of the interoperability of heterogeneous *networks*. A "datanet" could similarly solve the problem of the interoperability of heterogeneous data. It could enable both interoperability and unparalleled flexibility of extension.

"The time for big data warehouses is over," he said. "Data are big, and they are distributed." The amount of information they can provide goes beyond what humans can deal with. For example, his research group, which is looking at Huntington's Disease, can use 200 databases to create 120,000 potential predictions related to the disease. For this reason, he stressed the need for machine readability to find and use the ever-increasing amount of information available.

"Why aren't things changing?" he asked. As detailed in a recent EOSC report, 80 percent of the obstacles are social, not technical.[1] He highlighted what he called seven capital sins in the transition to open science: (1) older, more established scientists who tend to resist open science; (2) ignoring the complexity of data; (3) a gap between subject domain and infrastructure experts; (4) publication of data without supporting papers, and vice versa; (5) data presented in a way that machines cannot mine; (6) refusal to invest in research infrastructure; and (7) creation of data without a data stewardship plan.

Change is occurring. In Europe, FAIR principles, and what is known as the Internet of FAIR Data and Services, are gaining political traction, including a statement issued at the September 2016 G20 Summit to "promote open science and facilitate appropriate access to publicly funded research results."[2] In October 2017, the G7 issued a statement supporting the funding of the information infrastructure.[3] As a general guideline, an average of 5 percent in a research proposal budget should be dedicated to data stewardship, and every proposal should have a data stewardship plan. Better training for data stewards is also needed.

The Global Open (GO) FAIR will implement the recommendations of the EOSC, focusing on changes in culture, technology, and training, or GO Change, GO Build, and GO Train. "We hope to see the EOSC as the European contribution to a global Internet of FAIR Data and Services, and we are happy to work wherever possible with the United States," he concluded.

THE RESEARCH DATA ALLIANCE AND GLOBAL DATA INFRASTRUCTURE

Mark Parsons, Rensselaer Polytechnic Institute and former Secretary General of the Research Data Alliance (RDA), urged attention to two basic principles: first, to save, and second to share data. As a prerequisite, data must be preserved and safe; once saved, however, sharing data is harder to achieve because of a lack of trust and culture.

Infrastructure is about connections. Referring to a paper on the topic, he noted infrastructures can become "ubiquitous, accessible, reliable, and transparent as they mature."[4] They are comprised of relationships, interactions, and connections between people, technologies, and institutions. The RDA uses a bottom-up approach to build the

[1]European Commission. 2016. Realising the European Open Science Cloud: First Report and Recommendations of the Commission High Level Expert Group on the European Open Science Cloud. Online. Available at http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. Accessed November 28, 2017.

[2]European Commission. 2016. G20 Leaders' Communique Hangzhou Summit. Online. Available at http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm. Accessed December 15, 2017.

[3]G7 (Group of Seven). 2017. G7 Science Ministers' Communiqué. Turin, Italy, 27-28 September, 2017. Online. Available at http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqu%C3%A9.pdf. Accessed December 15, 2017.

[4]Edwards, P. N., S. J. Jackson, G. C. Bowker, and C. P. Knobel. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Online. Available at https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf?sequence=3&isAllowed=y. Accessed November 28, 2017.

social and technical connections, or bridges, that enable the sharing of data. Its 6,000 members vary in their interests and areas of focus, but common themes have emerged from their discussions: the need for persistent identifiers,[5] certifying trust, and recognition of the value of conversations, relationships, and mediation that create an agile network effect.

He asserted that "trust" crosses all three themes. It has social implications related to authority ("do I believe you?") and technological implications related to authenticity ("do I believe the object?"). RDA tries to facilitate interactions by providing a neutral forum, involving early adopters, and open problem-solving in all working groups and activities (see Figure 1).

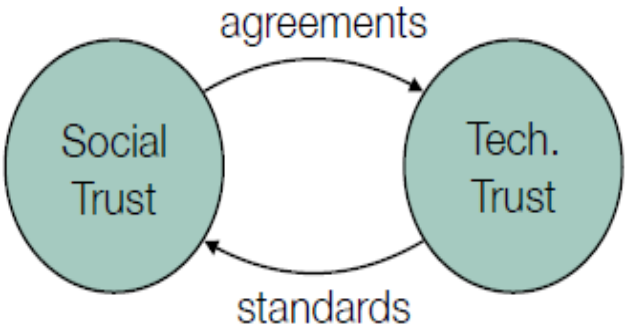

Figure 1 An ongoing interplay between social and technical trust.
Source: Mark Parsons, presentation, November 1, 2017, Washington, DC.

Parsons closed with six observations to achieve a global infrastructure: (1) working timelines should be kept short (12–18 months) to focus effort; (2) seed funding is a great help for adoption and deployment, but not for ongoing coordination, with a balance needed between central and local funding; (3) it is necessary to foster discussion fora and neutrality; (4) solid principles guide difficult decisions, which is where FAIR and the core principles of RDA help; (5) openness makes for more durable decisions; and (6) friction and disagreement based on different views (rather those rooted in power dynamics) are necessary and productive.

OPEN RESEARCH DATA POLICIES AND PRACTICES IN CHINA

To understand efforts in China to create an open data infrastructure, **Lili Zhang**, Chinese Academy of Sciences, provided an overview of several large projects. The National Science and Technology Infrastructure within the Ministry of Science and Technology supports 13 interdisciplinary data centers and platforms. In the last several years, evaluation mechanisms have developed to direct funding. The Chinese Academy of Sciences also promotes collection and sharing of data. More than 350 databases and 1,350 datasets in about 60 institutions have made available more than 600 terabytes of data for open access and downloading.[6]

Individual programs have developed policies related to data-sharing; a national-level policy is expected in 2018. Key policy players and research funders, in addition to the Ministry of Science and Technology and Chinese Academy of Sciences, include the China Association of Science and Technology and National Natural Science Foundation. Key promoters for open data include the National Science and Technology Infrastructure Center, Chinese Academy of Sciences, CODATA China and WDS China. A new program, CASEarth, has also begun to push data-driven discovery and decision-making.

A survey of Chinese scientists to understand barriers to data sharing yielded insights from about 7,000 respondents (see Figure 2). The two top barriers cited were uncertainty about intellectual property of the data and concern that publishing data would undermine competitive advantage. In descending order of frequency, other barriers included that the funder or journal did not require it, confidential or sensitive data were used, and lack of time or funding to publish them. Also cited were lack of incentives, uncertainty how to go about doing it, belief that data-sharing is not the researcher's responsibility, and privacy or ethical concerns.

Zhang said challenges and opportunities exist at different levels—from the overall policy, financial and social environment, to organizational issues, to incentives and other practical issues for scientists. New types of data publications may be a way forward.

INSIGHTS FROM BIOMEDICINE

George Komatsoulis, CancerLinQ, focused on the huge increase in amount and variety of data in a context of constricted resources. As an example in the growth and variety of data, the National Center for Biotechnology Information had 13.8 quadrillion bases of DNA sequence information in 2015—yet had none in 2009. Related to variety, Cancer-

[5]A persistent identifier is "a long-lasting reference to a digital resource. Typically it has two components: a unique identifier; and a service that locates the resource over time even when it's location changes," according to the Digital Preservation Coalition.
[6]For more information, see http://www.csdb.cn.

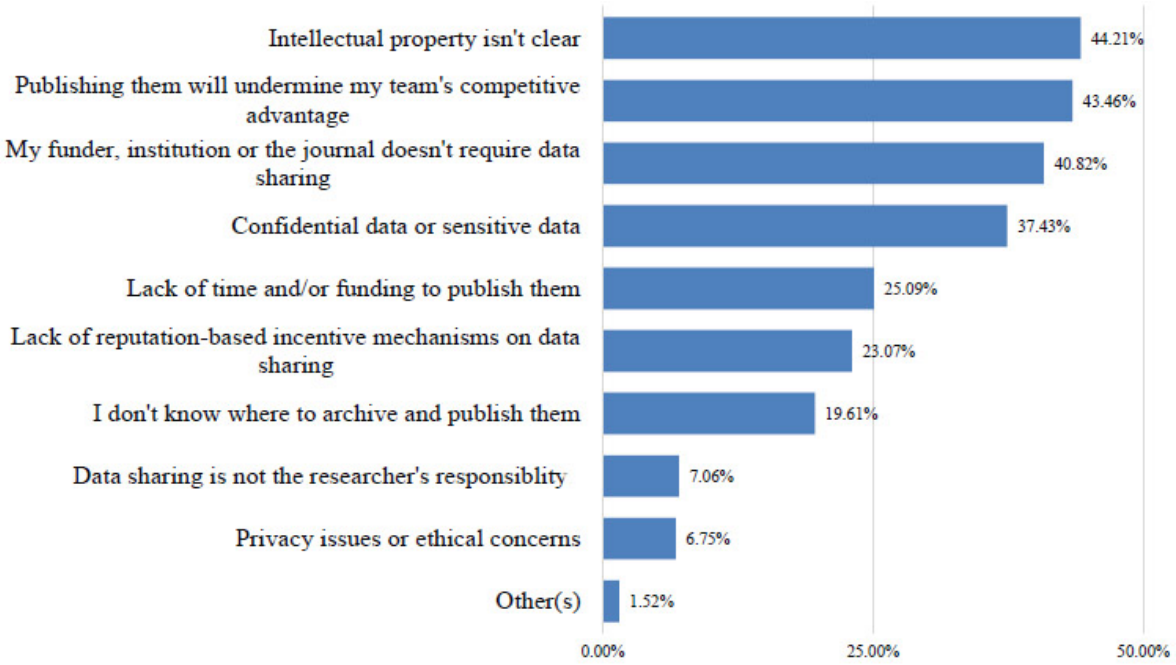

Figure 2 What are the barriers?
Source: Lili Zhang, presentation, November 1, 2017, Washington, DC.
Note: Reasons why respondents are hesitant to share their data, n=7082.

LinQ, a data-sharing program of the American Society of Clinical Oncology, pulls data from 710,000 patients that include diagnostics, laboratory tests, medications, and other structured and unstructured data. At the same time, research funding, as reflected in the National Institutes of Health (NIH) budget, has decreased. Thus, fewer resources are available to deal with more data.

The challenge is to improve return on investment by increasing value and reducing cost. Cost reductions include better efficiency related to archiving, storing, and managing data, as well as reducing redundancy. Increasing value can mean extracting more knowledge from each research effort. "It implies that the data doesn't become meaningless electrons, which is honestly the fate of a vast amount of the data that is out there," Komatsoulis said. He called for a data lifecycle, in which some data are maintained in real time at higher expense, while other data are archived at much lower expense to use as needed. FAIR principles can guide these decisions.

One way to manage data is through a Commons, such as the NIH Commons. Thus, instead of 15 copies of giant datasets, a single set can be accessed through the cloud. He described incentivizing good behavior to support a commons through a concept called Cloud Coins, which NIH is running as a pilot.[7] Other entities have expressed interest in the concept.

Echoing what others emphasized throughout the symposium, metadata is critical for interoperability. "You need to know what you are dealing with if you are going to use it in a scientifically meaningful way," he stressed. He called for standards, but sparingly, based on practical use, and as a way to help and not make things more difficult. The International Cancer Genomic Consortium, in which multiple countries are sharing data, and CancerLinQ, which is aggregating de-identified patient data to improve research, are two examples of these ideas in action.

NATIONAL INSTITUTES OF HEALTH ACTIVITIES

Symposium chair **James Hendler** highlighted NIH data-sharing activities in the absence of the scheduled speaker, who was unable to attend. Current challenges include the generation of large volumes of biomedical data, which are inexpensive to generate but costly to store on local servers. Multiple copies of the same data often reside in different locations. Some programs have built data resources that others cannot easily find.

Picking up from the previous presentation, NIH is developing a data commons to enable investigators to leverage all possible data and tools through collaborative research and robust engineering. All data are treated as digital objects that exist in a shared virtual space in compliance with FAIR principles. The commons allows transactions to occur on FAIR data at scale.

[7]For more information, see https://commonfund.nih.gov/bd2k/cloudcredits.

Considerable agreement exists about the general approaches to developing a data commons, and many of the technical problems are being addressed. Other challenges include community endorsement of best practices and standards in a rapidly evolving field. The NIH Data Commons Pilot is underway to allow access, use, and sharing of large, high-value biomedical data in the cloud.

EMERGING EFFORTS FOR SCIENCE DATA INFRASTRUCTURE

Amy Brand, MIT Press, moderated a panel on the National Science Foundation's (NSF's) data infrastructure efforts as well as two new initiatives: Metadata2020 and the Cross-continental and Management Pilot (C2CAMP).

Data and Knowledge as Infrastructure

Data and knowledge need infrastructure, stressed **Chaitan Baru**, NSF. "That sounds simple but implies a lot," he said, including researchers thinking about what they produce in a different way. "Getting to data should be easy, not a big burden," he added. The motivation behind setting up infrastructure is to facilitate easier and better access to data. The size of datasets and how to use the data, especially from multiple and heterogeneous sources, can be difficult. He also stated that there is a challenge to create interfaces that turn analytical findings into impactful data-driven stories for a diverse range of audiences.

One of NSF's six strategic directions—Harnessing the Data Revolution—has three research themes related to data sharing: foundational research into data science, systems and algorithms to exploit data and knowledge, and data-intensive research in all areas of science and engineering. Education and training are important. The commercial sector is using data at scale, and he called on the scientific community to do the same. "We want to be able to support integrative analysis and interpretation of multiple data and develop more advanced interfaces to...have dialogue-based interactions, and explanatory storytelling interfaces," he said. NSF has funded a number of projects related to creation of knowledge bases, creation of ontologies, knowledge extraction, knowledge aggregation, and reasoning.

In July 2016, a workshop on semantic information processing was held through the White House Office of Science and Technology Policy, bringing together industry, academia, and government. It led to two meetings in 2017 discussing the creation of an Open Knowledge Network. A report about next steps is in preparation.[8]

Metadata2020

L.K. Williams and **Scott Wymer**, Interfolio, described Metadata2020, which launched in September 2017 as a "collaboration that advocates richer, connected, and reusable metadata for all research outputs." Founded by Crossref and led by associations, publishers, and other supporters of scholarly data, it was set up to fuel discovery, connect metadata bridges, and eliminate duplication of effort.

Metadata 2020's Phase 1 involves listening to stakeholders that include publishers, librarians, funders, service providers, repositories, and researchers. Phase 2 will use this input to develop business cases and a Metadata Maturity Model against which content creators can measure themselves and improve. About 65 people have been involved in working groups to date. Williams and Wymer said more stakeholders are welcome to join the groups or provide input on benefits and challenges of a Metadata Maturity Model.

Cross-continental and Management Pilot (C2CAMP)

The Cross-continental and Management Pilot (C2CAMP) is a new initiative just getting underway according to **Larry Lannom**, Corporation for National Research Initiatives. C2CAMP involves several countries, including the United States, in a multiparty, distributed testbed based on open specifications. "The idea is to experiment with what are the minimal set of existing components and interfaces that can be put together and still deal with [digital] objects efficiently," he explained. Data producers and managers can prototype their work flows in the testbed.

"All of this data should be good news, resulting in higher levels of reuse, reproducibility, and accuracy," he commented. Instead, he said, science is dealing with a reproducibility crisis, funding issues, and excessive time spent on data management. C2CAMP is focused on making scientific workflows more efficient and on harmonizing workflows across domains. "Making data available without making it understandable may be worse than not making it available at all," he asserted. "It is then subject to misuse and misunderstanding." C2CAMP is a prototype of a distributed environment based on a digital object model. Everything in the environment is a digital object assigned a globally unique and actionable identifier, and is typed. "We want to start with a minimal set of components and services that

[8]For more information, see https://www.nitrd.gov/nitrdgroups/index.php?title=Open_Knowledge_Network.

enable the digital object model, identifiers, and a resolution system, then open the environment to as many use cases as possible to hone the core infrastructure," he explained. There is too much data for people to look at it all. Instead, data can be typed automatically, and the types will assist with workflow. C2CAMP will help determine how this can happen.

Discussion

In a discussion period after the presentations, one participant noted the importance of the symposium in sharing information about various initiatives. He made a strong plea to work internationally and in a community-driven fashion from the outset.

Another participant suggested an education campaign for funding agencies about persistent identifiers. "Funders think an identifier is when somebody stamped a number onto a PDF document," he commented. Thus, he said, they do not understand the need for investing millions of dollars in the process to make identifiers available and accessible. "Education for people who would fund these sorts of things might get us further along in terms of getting a usable, functional set of identifiers and an identification scheme that we can use," he suggested.

RESEARCH COMMUNITY PERSPECTIVES

The final panel of the symposium, moderated by **Sarah Nusser** of Iowa State University, looked at issues and opportunities arising from current and future data-sharing efforts.

Licensing Model and Ecosystem of Data Sharing

Jane Greenberg, Drexel University, pointed out that shared data in closed environments have played and will continue to play a role in science and other fields. Examples include the Inter-Collaborative Cancer Cloud, Collaborative Genomics Cloud, and the Fair Isaac Cooperation for consumer data. She discussed an NSF-funded project that she co-directs, the Licensing Model and Ecosystem for Data Sharing. Part of the Spokes Initiative and involving Drexel, MIT, and Brown University, it will develop a first-phase Knowledge Organization System for developing models for sharing restricted data and carrying out prototyping.

Barriers to sharing data in closed environments include policies, licensing, security, rights, privacy, and incentives. "The scenario that motivated our research is industry or government that has very secure or private information and wants to share their data with computer scientists to do some kind of predictive analysis, try some new algorithms, and so forth," she explained. Despite good intentions, it can take months to hammer out the legal and other agreements to share restricted data. The project is looking at developing a licensing model and ecosystem for data-sharing that can cover many situations to facilitate the process.

A recent workshop explored issues from the trenches to enable seamless data-sharing in industry and academia. Next steps are to collect data-sharing agreements from successful partnerships, build a trusted platform, and develop good metadata. She noted that collecting the agreements has been difficult and time-consuming, but the project has analyzed the approximately 30 agreements received to date. The goal is to develop a licensing framework with standard terms to which researchers, lawyers, and compliance teams can conform.

Accelerating Translational Medicine Using Heterogeneous Data: A Case for Better Metadata

Embracing heterogeneity rather than reducing it can address reproducibility problems and accelerate translation of basic science discoveries into clinical practice, suggested **Pervesh Khatri**, Stanford University. The traditional experimental approach of reducing heterogeneity tries to limit all variations when focusing on a disease. He presented an alternate framework that uses public-domain data to leverage heterogeneity. "What if we could account for heterogeneity from the beginning?" he posed. "What if we started with lots of data that is very heterogeneous...and then show that the hypothesis that we generated despite all these confounding factors is applicable in a broad spectrum of real-world patient population heterogeneity?"

The framework relies on discovery and validation of public-domain datasets, with a key concept that these datasets are considered separately. The framework has been used to diagnose sepsis 1 to 5 days before clinical suspicion, to predict response to an influenza vaccine, and to identify new drug targets, among other uses.

To expand the effort, the Center for Expanded Data Annotation and Retrieval (CEDAR) has been building a work bench to develop better metadata from the outset. Tools consist of a metadata authoring template, a template for authors' data, and a metadata repository. The reproducibility crisis is "a reporting reproducibility crisis, not data reproducibility," he posited. "In order to avoid this, we need better metadata." CEDAR's templates and other tools allow authors to more quickly select ontology and annotate data.

Automating Assessment of FAIR Project Architecture

"A scientific article is a research tool," said **Myles Axton**, Chief Editor of Nature Genetics. "The impact is from making available your method, your data, your code, your concepts, and your standards for the use explicitly of other people." Data can be put in limitless combinations with other datasets and code to generate unlimited knowledge. However, the frame in which timely queries can be aimed at the dataset is limited.

The field of research genetics provides a model for automating and incentivizing the sharing of data. The community has been willing to put code in GitHub and big data into public repositories in standard formats. Data reuse is also incentivized. Some of the most downloaded and cited papers in the field have analyzed existing data in public repositories, rather than creating new data.

Pilot experiments with the Layden group have attempted to automate standards writing and compliance checking for articles and other publications. In this context, a data steward helps domain-specific geneticists get their data in a form where other people's datasets can be automatically machine combined to generate new insights. Metadata modeling allows a researcher's data to work with other datasets. The same structures need to exist in different repositories so models can interact.

DARPA and Data: An Overview

The Defense Advanced Research Project Agency (DARPA) is a project-based agency that "dreams of data," according to **William Regli**, DARPA. Beginning in 2014, interest increased across the agency to "exploit data and computation to revolutionize or accelerate scientific discovery, innovation, and some engineering domains."

The agency stepped back to look holistically at current and future data needs. Computation will change in the future, thus affecting the costs and other factors associated with it today. Thus, one experiment is testing the hypothesis that rich metadata and standards are not needed, because "the tools are going to get better and better. Rather than burdening current performers, organizing their data for uses that they cannot imagine, what I would rather do is...preserve data in some form that we can make use of later," he said.

"It is really about organizing and optimizing the allocation of a human-machine team to answer questions from the data," he stressed. He framed this using the progression by which data evolves to information, then knowledge, understanding, and ultimately wisdom (see Figure 3). DARPA has made substantial investments in its portfolio to improve the computer-human interface, in which a machine is "not just showing me a pattern, but helping me to develop theories or abstractions that can be used to take action," he stated, and elevating the computer from a tool to a partner in this evolution.

Discussion

In a discussion period after the presentations, education emerged as a topic. One participant noted the importance of educating lawyers involved in data sharing. Another commented that data science programs need to train the next generation of data stewards. "One size does not fit all," the participant commented. "We need to be thinking about the skillsets of the people doing this work."

Another participant noted the value of distributive sharing of data in which a workflow can visit a researcher's data but the researcher does not cede control of it. "I think distributive learning over data that do not leave my 'container' is the future, not only because of costs but also because of the psychological resistance about giving data to anyowner," he said. Another participant agreed with the concept of a distributed approach but warned against giving the original data producer too much control. "The last person to ask about data reuse is the person who created it," he said. "They do not understand how it could be reused."

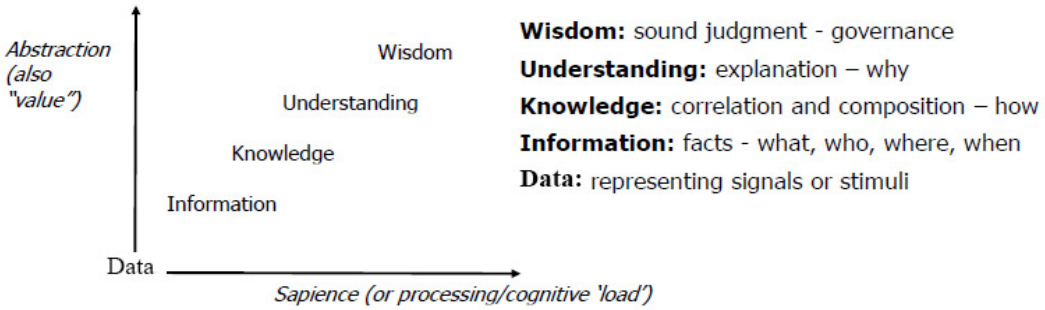

Figure 3 Working toward wisdom.
Source: William Regli, presentation, November 1, 2017, Washington, DC.

CONCLUDING REMARKS

Symposium chair **James Hendler** closed by reflecting on points he heard throughout the day. He said, one of the key things that we have heard is that the FAIR principles are winning acceptance. Efforts must now focus on operationalizing the principles for humans and for machines. "People have to want to share their data, understand how to share the data, and want to reuse the data," he said. "Machines have to be able to take and process the data."

There is a need to share models, best practices, and lessons learned. Different fields of science have different cultures, so one-size-fits-all will not work. Astronomy has had a culture of sharing, for example, in part because of limited access to the equipment to conduct observations and experiments. Other fields of science and engineering have less tendency to share data. Infrastructure to span disciplines must span different cultures. Technological change, in which researchers see the benefits gained by sharing data, may drive cultural change.

As a community, Hendler urged researchers not to look at one solution, but rather at interoperability between the various approaches being tested. Drawing an analogy from development of the Internet over the past few decades, the World Wide Web beat competitors as the dominant presence because it is a simple standard and created interoperability. The lesson for data-sharing is not to make things too complicated. "Find commonality and grow on top of that," he noted as a common theme expressed throughout the symposium.

He reflected the recurring point that metadata is crucial to discover, integrate, validate, and draw explanations from the thousands of datasets around the world. Many participants also noted that funding agencies are an important driver in moving ahead with data sharing.

DISCLAIMER: This Proceedings of a Workshop—in Brief was prepared by **Paula Whitacre** as a factual summary of what occurred at the meeting. The statements made are those of the rapporteur or individual meeting participants and do not necessarily represent the views of all meeting participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine.

REVIEWERS: To ensure that it meets institutional standards for quality and objectivity, this Proceedings of a Workshop—in Brief was reviewed in draft form by **Amy Brand**, MIT Press and **Peter Wittenburg**, Max Planck Computing and Data Facility. The review comments and draft manuscript remain confidential to protect the integrity of the process.

PLANNING COMMITTEE: James Hendler, Rensselaer Polytechnic Institute; **Larry Lannom**, Corporation for National Research Initiatives; **Barend Mons**, Leiden University Medical Centre; and **Sarah Nusser**, Iowa State University. Staff: **George Strawn**, director, Board on Research Data and Information (BRDI); **Ester Sztein**, deputy director, BRDI; **Emi Kameyama**, associate program officer, BRDI; and **Nicole Lehmer**, senior program assistant, BRDI.

SPONSORS: This symposium was supported by the National Science Foundation and the National Institutes of Health as part of BRDI activities, and the publication of this Proceedings of a Workshop—in Brief was supported by the National Academies of Sciences, Engineering, and Medicine.

For additional information regarding the workshop, visit www.nas.edu/brdi.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2018. *International Coordination for Science Data Infrastructure: Proceedings of a Workshop—in Brief.* Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25015.

Board on Research Data and Information
Policy and Global Affairs

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org