OMICS A Journal of Integrative Biology
Volume 15, Number 4, 2011

DOI: 10.1089/omi.2011.0007

Policy and Data-Intensive Scientific Discovery in the Beginning of the 21st Century

Vural Ozdemir,[1] Charles Smith,[2,3] Kathleen Bongiovanni,[2] David Cullen,[2] Bartha M. Knoppers,[1] Andrew Lowe,[2] Mette Peters,[3] Robert Robbins,[4] Elizabeth Stewart,[2] Gene Yee,[5] Yi-Kuo Yu,[6] and Eugene Kolker[2,3]

Abstract

Recent developments in our ability to capture, curate, and analyze data, the field of data-intensive science (DIS), have indeed made these interesting and challenging times for scientific practice as well as policy making in real time. We are confronted with immense datasets that challenge our ability to pool, transfer, analyze, or interpret scientific observations. We have more data available than ever before, yet more questions to be answered as well, and no clear path to answer them. We are excited by the potential for science-based solutions to humankind's problems, yet stymied by the limitations of our current cyberinfrastructure and existing public policies. Importantly, DIS signals a transformation of the hypothesis-driven tradition of science ("first hypothesize, then experiment") to one that is typified by "first experiment, then hypothesize" mode of discovery. Another hallmark of DIS is that it amasses data that are public goods (i.e., creates a "commons") that can further be creatively mined for various applications in different sectors. As such, this calls for a science policy vision that is long term. We herein reflect on how best to approach to policy making at this critical inflection point when DIS applications are being diversified in agriculture, ecology, marine biology, and environmental research internationally. This article outlines the key policy issues and gaps that emerged from the multidisciplinary discussions at the NSF-funded DIS workshop held at the Seattle Children's Research Institute in Seattle, on September 19–20, 2010.

Introduction

Systematic approaches to understanding the future of innovations and the development of attendant science policy predate to the time of Great Depression in the United States (U.S.). This was an era when fundamental links between technologies, social, and economic development were actively explored. Notably, the sociologist William F. Ogburn examined the patterns of innovations at that time (Ogburn, 1922). Ogburn developed the concept of "cultural lags," referring to the need for societal adaptations to technology, and that the society lags in response to technological advances. Implicit in Ogburn's cultural lags approach to inventions was a *linear* model of innovations whereby technological advances in science would subsequently impact the society (Godin, 2006).

Science policies in 20th century were reactive, often triggered as a response to crisis in practice of science. This is not an optimal approach to policy making because reactive policies inherently resort to "playing catchup" or "damage control" from unintended or detrimental effects of technologies that have already materialized. It can also result in loss of public trust in science that can take lengthy periods or even generations to remedy; e.g., consider the public discourses and controversies over genetically modified organisms, GMOs.

In contrast to the linear model of innovations that prevailed for the most part of the 20th century, we have recently seen a shift, however, toward conceptualization of innovations as nonlinear ecosystems with many moving parts that intersect and interact in various ways (e.g., competition, cooperation, precompetitive collaboration). Increasingly, we have to deal with the challenge of exponential growth in data volume, generated daily through massively parallel study of biological pathways in living matter or complex systems in engineering and natural sciences.

[1]McGill University, Montreal, QC, Canada.
[2]Seattle Children's Research Institute, Seattle, Wahsington.
[3]University of Washington, Seattle, Washington.
[4]Electronic Scholarly Publishing Project, Woodinville, Washington.
[5]Fenwick and West, Seattle, Washington.
[6]National Center for Biotechnology Information, NIH, Bethesda, Maryland.

Together with large volumes of data, crosscutting collaborations are becoming essential to enable discoveries in data-intensive "big science" projects such as the Human Genome Project in the health sector and the Sloan Digital Sky Survey ("Cosmic Genome Project") in astronomy. Beyond traditional peer-to-peer academic collaborations, cooperation among "extended community of peers"—networks of networks—that span governments, academia, industry, and various end-users of knowledge in society has become a *sine qua non* for 21st century science. Indeed, data-intensive science was named as the *Fourth Paradigm of Science*, preceded by the third (last few decades: *computational* branch, modeling, and simulating complex phenomena), the second (last few hundred years: *theoretical* branch, using models leading to generalizations), and the first paradigm (a thousand years ago: *empirical description* of natural phenomena) (Hey et al., 2009).

Importantly, this data-intensive Fourth Paradigm of science signals a transformation of the hypothesis-driven tradition of science ("first hypothesize-then-experiment") to one that is typified by "first experiment-then-hypothesize" mode of discovery. Open access to large volumes of data is therefore a key prerequisite for discoveries in the beginning of the 21st century science. Yet the data-centric mode of modern scientific discovery and the requirement for open access to a data "commons" create hitherto unprecedented unique policy needs.

As DIS innovations are now rapidly diffusing to applications not only in developed but also in low- and middle-income countries (LMICs) in the broader context of globalization of science (van Kerkhoff and Szlezák, 2006), we herein pose the following question: Can we do better in policy making for DIS, at this critical inflection point when its applications are being diversified not only in health but also in agriculture, ecology, marine biology, and environmental research internationally? Further, as DIS applications emerge in the broader context of 21st century science that uniquely emphasizes the integration of life sciences and the humanities (UNESCO, 2000), what are the unique "human factors" and nuanced social issues that need to be incorporated with technical factors in DIS policy making?

We underscore that policies are crucial tools for system level analysis and facilitation of DIS innovations rather than piecemeal progression of DIS with an ad hoc policy infrastructure. Significance of the need for DIS policy becomes clearer as this form of high-throughput science involves not only biological systems but also the environment(s) living organisms are embedded in. As such, proliferation of DIS data is a result of both high-throughput technology and intensive analysis of the environmental variables that interact with the host. A corollary of this vision is that DIS is firmly embedded in a social or societal context as a prototype example of 21st century science. If we are to address the challenges and effectively reap the benefits of DIS in 21st century, DIS policies are needed that consider both technical and social barriers in real-time, and transform them into opportunities for sustainable growth of DIS (Guston and Sarewitz, 2002; Selin, 2008).

Finally, the present DIS policy and foresight working group recognizes that policies are inherently *living* documents and that policy making should ideally span from cell to society and public policy in real time with scientific advances, and remain responsive to anticipated future trajectories of DIS. Hence, the current DIS policy and foresight report is framed around four subchapters from (1) current state of DIS policy, (2) barriers, and (3) future outlook to (4) representation and engagement with multiple publics. The latter is referred to as "society" though this includes all conceivable stakeholders both expert, lay, and others engaged in creative generation or use of DIS data and knowledge.

The policy gaps and analyses listed below directly inform the recent *NSF Cyber Infrastructure Vision* that rests on the "the development of a cultural community that supports peer-to-peer collaboration and new modes of education based upon broad and open access to leadership computing; data and information resources; online instruments and observatories; and visualization and collaboration services" (Bement, 2007). The workshop report comes at a time when the NSF Office of Cyberinfrastructure Task Force on Data and Visualization (NSF, 2011) is also making its recommendations, and the authors find that many issues discussed at the workshop are echoed in that report.

Current State

Current policies do not support the DIS to a level commensurate with the fourth paradigm in science practice and governance (Hey et al. 2009). A large body of policy efforts exist (e.g., Minimum Information About a Microarray Experiment MIAME Standards) at the level of *analytical validity* (e.g., data capture, analysis) but much less on ways to translate the DIS data to value-added products, whether they be in the health sector, agriculture, or ecology. It is also not clear whether and to what extent the existing policies on DIS data analytical validity had an impact on DIS data quality. More empirical research on existing and future DIS policies is essential for continued vigilance on the appropriateness of the policy-making process for DIS.

A hallmark of DIS is that it amasses data that are public goods (i.e., creates a "commons") that can further be creatively mined for various applications in different sectors. As such, this calls for a science policy vision that is long term. Yet most research funding agencies maintain a short-term approach, which may or may not sustain the translation of DIS into tangible value added products. Policy measures also need to protect the "DIS commons" so that fractures in this commons are prevented, a concern that is not unrealistic given the current fragmentation of DIS policies that are regrettably discontinuous at the level of clinical validity and utility.

Forging linkages between different standards communities will be essential to enable discoveries in data-intensive sciences. Large volumes of data are captured on phenotypic variability in living matter (e.g., disease susceptibility, responses to drug and nutritional exposures or infectious agents, plant-related traits in agriculture). Discoveries are often made by systematic associations between such large phenotypic and biological data sets. Data-intensive science policy should also consider how best to streamline the standards on both phenotypes and various large-scale biological datasets.

We suggest that there is a need for effective linkage (and convergence) of the data standards community with those engaged in knowledge translation, public health, and policy. The Canadian Institutes of Health Research (CIHR), the largest funding body for health research in the country, notes

that knowledge translation "involves an active exchange of information between the researchers who create new knowledge and those who use it" in its report: Knowledge Translation Strategy 2004–2009 (Knowledge Translation Working Group, 2004). The CIHR underscores that knowledge translation is "radically different from the traditional view of 'knowledge transfer' as a unidirectional flow of knowledge from researchers to users. In this traditional model, not surprisingly, low success in knowledge uptake was attributed to the 'two communities' problem in which researchers and policy makers inhabit different worlds with different language and culture." For effective knowledge translation, bringing together creators and users of data, as well as of standards during all stages of the research cycle, is essential.

Biological complexity captured by high-throughput data translates into knowledge after robust association analyses with clinical phenotypes (e.g., disease susceptibility, responses to drug and nutritional interventions) and public health outcomes. Notably, in the case of genetics/genomics data, measures to strengthen the reporting of genetic association analyses have been recently taken by the STREGA (STrengthening the REporting of Genetic Association studies) recommendations. The STREGA Statement proposes a minimum checklist of items for reporting genetic association studies. However, the STREGA recommendations do not prescribe or dictate how a genetic association study should be designed, but seek to enhance the transparency of its reporting, regardless of choices made during design, conduct, or analysis. Current efforts for omics data standards (reporting, sharing, etc.) would be well served by further engagement with the above initiatives in clinical investigation and public health communities who utilize omics data. Specifically, establishing effective linkages between the ongoing minimum information checklist development projects in the omics knowledge domain with initiatives such as STREGA, might be actionable concrete next steps for broader knowledge translation. This also means, however, that the concept of "community standards" in DIS needs to be revisited so that a broader range of stakeholders are included in defining the "community" from upstream discovery DIS and association analyses to public health and policy knowledge domains.

Barriers

Policies have been mostly in the context of analytical validity but considerably less in the context of clinical validity, utility, or the social and ethical contexts. There is a need to connect the policy communities who often work separately in data capture, analysis, and quality with those engaged in translating the DIS data to various applications and value-added products in public health, agriculture, and ecology or global health.

Given that DIS is highly collaborative in nature, policies that support new indicators (e.g., bibliometric measures other than first or senior authored publications) of individual contributions to collective work need to be developed. Still, many of the academic promotion committees or other types of scientific recognition contexts heavily rely on individual contributions or "discovery of a certain molecule or well encapsulated body of work"—in the face of a DIS that is inherently based on collective work.

Such human and social factors are too significant to neglect in the Fourth Dimension that the 21st century science is increasingly embedded in. Taken together, the reward systems in scientific practice should be appropriately modified for DIS to move forward with genuine contributions both from individuals and DIS networks. A related issue is that the service component that is so essential to amass large DIS data sets should be recognized with appropriate reward or incentive mechanisms—that is, why not establish postdoctoral fellow, graduate student, or technician awards to recognize such service work that further creative analysis badly depends on? Why not have journals require that data generators be recognized as authors or contributors?

DIS in part rests on establishing mechanism-based associations (i.e., a correlative science) between DIS biology data and various phenotypes such as susceptibility to common complex diseases, individual responses to drugs and nutritional exposures, or radiation and infectious agents, to name a few. As such, two crucial pillars of the DIS are both high-throughput biological data and phenotype. Any discussion on DIS policy should then capture these two dimensions of DIS. Interestingly, although the efforts on DIS policy have been made for biological components of DIS, policies on how best to sustain capture, analysis, and interpretation of *phenotypic* datasets remain poorly developed.

Errors might conceivably be larger in DIS because errors at the level of a singular gene or phenotype would accrue and exponentially proliferate when DIS expands the scope of scientific inquiry to genomes and phenomes (i.e., study of all plausible and measurable phenotypes), respectively.

Although the use of phenotypic data from routine healthcare services might allow optimization and reduce redundancy of DIS resource utilization, this also can create a tension between phenotypic data that is generated primarily for a healthcare service goal versus primarily a research driven agenda. In other words, phenotypes that are not measured with research in mind can be shaped by other social and human factors such as economics or administrative confounding. For example, phenotype data that are recorded to meet the demands and practices of administrative priorities will not always be suitable for DIS association analysis with high-throughput DIS biology data. One potential remedy could be to capture the phenotype data in a raw form (i.e., before it is codified in any administrative or health service oriented context) such that its use in DIS can be facilitated among the user groups. Nonetheless, policies on the DIS phenotype datasets remain a challenge as well as how best to obtain DIS phenotype data where service and research mandates can compete and conflict with data quality.

Finally, how much flexibility can or should be tolerated in DIS policy making, given that DIS itself is a heterogeneous science (genomics, proteomics, metagenomics, astronomy, etc.) generating heterogeneous technical, social, cultural issues—each of which might require nuanced and perhaps customized policy options?

Future/Outlook

DIS promises discoveries for unprecedented mechanisms that underpin common diseases or response to environmental exposures, agriculture-related traits, and ecology. This vision is firmly embedded in a theme of expansion of science and data.

But in the history of science, it is not uncommon to witness periods of expansion and contraction. During the Renaissance, many recognized artists were also engineers with crosscutting perspectives. Hyperspecialization in the later part of 20th century resulted in the creation of artificial knowledge silos. For future DIS policies to be sustainable, we need to be prepared for both DIS as well as a form of science that is perhaps less data-intensive. Put simply, we need policies that can address the needs of both hypothesis free (DIS) and hypothesis driven science. Such a strategy would best prepare the DIS community for any eventual and unforeseeable expansions and contractions in the scope of 21st century science. A corollary of this proposal is that we need policies that support not only bench-to-bedside but also bedside-to-bench research. Hypotheses waiting to be discovered at the bedside—at the phenotypic outcome level—ultimately "trigger" fundamental mechanism-based bench research. In DIS policy making we cannot afford to neglect the wisdom of lessons from the bedside. Yet these lessons need not be solely provided by clinicians—PhD scientists also need policies that support their "visit" to bedside so that they can make observations through the lens of mechanism-based fundamental research, collaboratively with clinicians.

DIS would be well served in the long run by encouraging editorial and funding agency policies to allow individual investigators and authors of scientific manuscripts to think beyond the immediate implications of their own research findings or research proposal. This is much needed to cultivate a 21st century DIS culture that promotes lateral thinking whereby investigators and authors need to genuinely reflect on the broader significance of science and its impacts on society, and vice versa (science and society are coconstructed). Only with this bidirectional recognition, can we ensure a sustainable supply of Renaissance scientists that share a collective vision of science. An "Office of Broader Science Impact Analysis" would be timely to put these ideas into practice in the near to midterm future. Extending this idea further, an "Office of Ombudsperson for Prospective Foresight in DIS" would serve well to anticipate the future trajectories in DIS—while the future is yet undecided, and before ideas by each stakeholder are "locked" into deterministic futures.

DIS policy should firmly consider the emerging field of global health (Pang et al., 2010)—especially because of the 90-10 gap—90% of the World's health research funds are dedicated to health concerns that affect 10% of the global population. This gap could conceivably widen further with DIS if capacity is not developed for DIS in LMICs.

Policies that support filters to accept or reject DIS data and information are needed so that aggregate information can be rapidly generated from submitted DIS data.

DIS is essentially an enabling science that contributes to translational research not only in the health sector but also in other fields such as agriculture and ecology. DIS policy needs to work through these diverse applications, keeping in mind that each application context for DIS may raise nuanced and customized policy measures. Theragnostic applications of DIS (pharmacogenomics, nutirgenomics, etc.) require policy measures that deal with DIS applications to such emerging health interventions.

Representation and Engagement with Society

Innovation and scientific practice are often portrayed as the works of a lone genius operating in a laboratory. Although this might have been partially true in the early part of the 20th century, this vision is no longer accurate nor sustainable with the arrival and challenges of DIS. High-throughput science and innovations are no longer the products of a singular person or stakeholder. This demands a renewed vision on DIS policy that firmly recognizes that the key to a sustainable developmental trajectory rests in not only technological factors but also in societal issues and how science and technology shape, and are shaped by society.

Lessons from science and society interactions such as GMOs illustrate the need for prospective policy making in DIS. A sit-and-wait approach is not tenable. As noted in the introduction section, although science and technology may eventually find their appropriate trajectory in society, an ad hoc application of science can breach public trust or result in unintended consequences some of which may require generations (of publics) to rebuild trust and meaningful dialogue among stakeholders.

If DIS innovations resemble ecological networks with many visible (and often not so visible) interdependencies among stakeholders, how can we engage with and include different stakeholders? Even if we succeed in engagement of stakeholders in a context of innovation ecosystem, this tends to be carried out after opinions or firm value systems and expectations were already developed among the key stakeholders. This is problematic as once strong opinions are established it is difficult to establish an open dialogue or to shape the anticipated future trajectories by each stakeholder. Human behavior and opinions are not always easy to modify even in the face of strong scientific evidence.

We need to bear in mind that the goal of effective policies is not simply "regulation" or playing the role of an auditor but to support sustainable growth of DIS innovations and establishing a stakeholder negotiation platform wherein the future trajectories can be actively negotiated in the spirit of a participatory democracy among the stakeholders.

The NSF_OCI_TFDV report recommends "specific budget provision for the establishment and maintenances of data sets/services and the associated software and visualization tools infrastructure." Some enabling tools akin to Wikipedia can presumably be developed to put this vision into practice. Even though ideas on a DIS innovation trajectory may be initiated by experts, this can be further shaped by various publics through a "Wiki-innovation" type approach. In other cases it is conceivable that the end-users of DIS data and knowledge can launch an innovation wiki-platform and negotiation among the stakeholders.

Given that concerns have already been expressed to avoid the "two communities thesis" wherein experts and policymakers are detached in the flow and exchange of scientific information and knowledge, such a wiki innovation platform might perhaps bring previously isolated stakeholder communities closer.

Conclusion

Since Vannevar Bush (first modern science policy: *Science: The Endless Frontier*, 1945), more than 6 decades of inquiry reveal the scope and depth of complexity in biological and

natural systems. This remains a key challenge of modern science at the beginning of the 21st century. In fact, the need for systems approaches to study complexity has been recognized early on by physiologists who studied homeostasis at the former turn of the century in the 1900s. Systems approaches to study of innovations (and of the society impacted by innovations) is, however, a relatively more recent concept. In much the same way we cannot understand biological complexity by studying genes and proteins one at a time, innovations in 21st century too need to be understood as complex ecosystems with many moving parts (e.g., knowledge creators and users) that intersect and interact in a nonlinear fashion as noted earlier.

Lavis et al. (2002) proposed that "researchers (and research funders) should create more opportunities for interactions with the potential users of their research. They should consider such activities as part of the 'real' world of research, not a superfluous add-on.". On the other hand, efforts for standards development in omics sciences, by and large, have focused on the "data" (whether it is reporting or sharing) thus far. But omics data feed into and inform a much larger array of innovations in the beginning of the 21st century such as genotype–phenotype associations in clinical investigation and public health sciences (Knoppers et al., 2010); these fundamentally translate omics data into omics knowledge. If we continue to rely solely on the linear model of innovations, there is no guarantee that what has been (and will be) developed within the omics data standards community will effectively reach and have active uptake in clinical and public health knowledge domains.

To the extent that the goal of DIS is the creation of knowledge (i.e., not only high-throughput data per se) that meaningfully impact healthcare, public health and policy, novel knowledge translation platforms, and cyber infrastructure are essential to move research evidence to practice across the entire knowledge domains that make up the DIS ecosystem.

Acknowledgments

This policy report and DIS workshop were supported by SCRI and NSF Grant DBI-0969929 to E. Kolker (Principal investigator). V. Ozdemir is supported by an investigator salary award for science-in-society research on omics technology applications from FRSQ (Montreal, Canada). The views expressed in this article are entirely personal opinions of the authors and do not necessarily represent positions of their affiliated institutions or NSF.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Bement, A.L. (2007). Letter from the Director. National Science Foundation Cyber Infrastructure Vision for 21st Century Discovery. Available at: http://www.nsf.gov/pubs/2007/nsf0728/nsf0728_1.pdf (accessed February 3, 2011).

Bush, V. (1945). *Science: The Endless Frontier.* Washington, DC: U.S. Government Printing Office.

Godin, B. (2006) The linear model of innovation. The historical construction of an analytical framework. Sci Technol Hum Values 31, 639–667.

Guston, D.H., and Sarewitz, D. (2002). Real-time technology assessment. Technol Soc 24, 93–109.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, WA: Microsoft Research.

Knoppers, B.M., Leroux, T., Doucet, H., Godard, B., Laberge, C., Stanton-Jean, M., et al. (2010). Framing genomics, public health research and policy: points to consider. Public Health Genomics 13, 224–234.

Knowledge Translation Working Group. (2004). *Innovation in Action: Knowledge Translation Strategy 2004–2009.* Ottawa, Canada: Canadian Institutes of Health Research.

Lavis, J., Ross, S., and Hurley, J. (2002). Examining the role of health services research in public policymaking. Milbank Q 80, 125–154.

National Science Foundation, NSF (2011). Office of Cyberinfrastructure, Task Force on Data and Visualization (in press).

Ogburn, W.F. (1922). *Social Change.* New York: Dell.

Pang, T., Daulaire, N., Keusch, G., Leke, R., Piot, P., Reddy S., et al. (2010). The new age of global health governance holds promise. Nat Med 16:1181.

Selin, C. (2008). Sociology of the future: tracing stories of technology and time. Sociol Compass 2, 1875–1895.

UNESCO (Paris). (2000). World Conference on Science. "Science for the twenty-first century: a new commitment" (Declaration on Science and the Use of Scientific Knowledge). Sci Technol Soc 5, 81–92.

van Kerkhoff, L., and Szlezák, N. (2006). Linking local knowledge with global action: examining the Global Fund to Fight AIDS, tuberculosis and malaria through a knowledge system lens. Bull World Health Organ 84, 629–635.

Address correspondence to:
Vural Ozdemir, M.D., Ph.D.
Centre of Genomics and Policy
Department of Human Genetics
Faculty of Medicine
McGill University
Montreal, QC, Canada

E-mail: vural.ozdemir@mcgill.ca

OR

Eugene Kolker, Ph.D.
Seattle Children's Research Institute
1900 Ninth Avenue
C9S-9
Seattle, WA 98101

E-mail: eugene.kolker@seattlechildrens.org

This article has been cited by:

1. Fethi A. Rabhi, Lawrence Yao, Adnene Guabtni. 2012. ADAGE: a framework for supporting user-driven ad-hoc data analysis processes. *Computing* **94**:6, 489-519. [CrossRef]
2. Eugene Kolker , Elizabeth Stewart , Vural Ozdemir . 2012. Opportunities and Challenges for the Life Sciences Community. *OMICS: A Journal of Integrative Biology* **16**:3, 138-147. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
3. Vural Ozdemir , Samer A. Faraj , Bartha M. Knoppers . 2011. Steering Vaccinomics Innovations with Anticipatory Governance and Participatory Foresight. *OMICS: A Journal of Integrative Biology* **15**:9, 637-646. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]