Workshop Report

The Northeast Big Data Innovation Hub: "Enabling Seamless Data Sharing in Industry and Academia" Workshop

Drexel University, Philadelphia, PA, September 29-30, 2016
Report completed March 31, 2017

Written by:

Jane Greenberg, Alice B. Kroeger Professor, Director of the Metadata Research Center, College of Computing & Informatics, Drexel University

Samantha Grabus, Research Assistant, Metadata Research Center

Florence Hudson, Senior Vice President and Chief Innovation Officer, Internet2

Tim Kraska, Assistant Professor, Department of Computer Science, Brown University

Samuel Madden, Professor, Electrical Engineering & Computer Science, Massachusetts Institute of Technology, and Co-Director of the Intel Science and Technology Center (ISTC) in Big Data

René Bastón, Executive Director, Northeast Big Data Innovation Hub

Katie Naum, Program Coordinator, Northeast Big Data Innovation Hub

Workshop Sponsors:

Northeast Big Data Innovation Hub; the Computing Community Consortium; the Metadata Research Center, College of Computing and Informatics, Drexel University; and the Gerri C. LeBow College of Business, Drexel University.

Table of Contents

Executive Summary

Increasingly, both industry and academia, in fields ranging from biology and social sciences to computing and engineering, are driven by data (Provost & Fawcett, 2013; Wixom, et al, 2014); and both commercial success and academic impact are dependent on having access to data. Many organizations collecting data lack the expertise required to process it (Hazen, et al, 2014), and, thus, pursue data sharing with researchers who can extract more value from data they own. For example, a bio-sciences company may benefit from a specific analysis technique a researcher has developed. At the same time, researchers are always on the search for real-world data sets to demonstrate the effectiveness of their methods. Unfortunately, many data sharing attempts fail, for reasons ranging from legal restrictions on how data can be used—to privacy policies, different cultural norms, and technological barriers. In fact, many data sharing partnerships that are vital to addressing pressing societal challenges in cities, health, energy, and the environment are not being pursued due to such obstacles.

Addressing these data sharing challenges requires open, supportive dialogue across many sectors, including technology, policy, industry, and academia. Further, there is a crucial need for well-defined agreements that can be shared among key stakeholders, including researchers, technologists, legal representatives, and technology transfer officers. The Northeast Big Data Innovation Hub (NEBDIH) took an important step in this area with the recent "Enabling Seamless Data Sharing in Industry and Academia" workshop, held at Drexel University September 29-30, 2016. The workshop brought together representatives from these critical stakeholder communities to launch a national dialogue on challenges and opportunities in this complex space. Over the course of two days:

- Representatives from industry, government, academia, and nonprofit organizations shared case studies of real life data sharing successes, providing evidence of best practices that can be replicated or extended to address the range of legal, policy-based, privacy, technical, and other noted challenges.
- Representatives from stakeholder communities also shared examples of data sharing failures, including assessments of what went wrong and why. The two most prevalent factors contributing to data sharing failure included: **1)** legal concerns and implications, and **2)** the extensive time required to ensure a data sharing success. (With the latter factor, one case study presented clear evidence of a data sharing agreement, that while eventually successful, unfortunately came too late in the research process to be of any value).
- Participants collectively identified the most significant classes of data not being shared as follows: Genomics data, financial data, city planning data, and utility

data relevant to disaster preparedness and response activities. These types of data were highlighted and referenced multiple times throughout the workshops, in dialogue and as part of cases where data sharing obstacles are interfering with research and the pursuit of scientific and societal advancement.
- As a final task, participants scoped a list of potential solutions for addressing challenges, which include: **1)** standardized data-sharing agreements that could be used across organizations and for a range of partnerships; **2)** metadata standards and ontologies to classify the nature of data being shared and the type of data sharing agreement; and **3)** technological platforms that automate access control, revocation, anonymization, auditing, and hosting of data, while enforcing standardized data sharing contract terms.

The workshop serves as a first step of the NEBDIH's Data Sharing Working Group's efforts to engage participants and lay groundwork for collectively addressing data sharing obstacles extending beyond the open data community. The workshop activity was also interconnected with the recently announced Northeast Spoke award, "A Licensing Model and Ecosystem for Data Sharing" (PI Madden, co-PIs Binnig, Greenberg, Kraska, Weitzner). The chief aim of this new Spoke award is to develop a safe and secure platform that facilitates the sharing of data that may or may not be open or free between different organizations (industry, academia, government). To this end, the workshop successfully engaged industry, government, nonprofit, and research community stakeholders in shaping the NEBDIH's data sharing discussion, identifying challenges and opportunities to accelerate data sharing, and considering initial platform design goals integral to the Spokes award work plan.

Key outcomes of the workshop included **1)** a set of first-phase action items, and **2)** a list of long-term goals to help build a community-driven approach to addressing data-sharing problems. These outcomes are:

First phase action items:
- Identify a platform for continued dialogue and other collaboration opportunities, such as webinars or calls to share progress and obtain feedback from the larger ecosystem. Participants expressed desire for continued engagement with the NEBDIH and across the hubs (Slack and an open wiki were both suggested as options).
- Interface with select, impactful communities (e.g., Smart Grid or Precision Agriculture) that are eager to share data and interested in collaborating to develop solutions.
- Develop a toolkit of resources that can both inform stakeholders about existing challenges, and provide access and pointers to solutions, including developments underway.

Long-term goals:

- Establish both a method and a system for people to communicate, via a common language, what their data agreements need to convey. The method and system should provide lawyers attorneys and other constituents with the necessary legal language, following principles that lead to successful data sharing agreements.
- Confirm that the method and system developed to address data sharing barriers supports the efficient minting of a variety of data sharing agreements, following a menu of policies with regards to privacy, PII (personal identifying information), future data sharing and repurposing, and other considerations.
- Create a mechanism that allows organizations to conduct efficient and effective risk analysis for specific data sharing agreements.

Organization and Participant Representation

The "A Licensing Model and Ecosystem for Data Sharing" workshop was a two-day event, running September 29-30, 2016. The workshop was held at Drexel University in Philadelphia, Pennsylvania, and sponsored by the Computing Community Consortium. Additional support was provided by the NEBDIH; the Metadata Research Center, College of Computing and Informatics, Drexel University; and the Gerri C. LeBow College of Business, Drexel University.

Workshop content and planning was overseen by a coordinating team that included: Jane Greenberg, Professor, Director of the Metadata Research Center, College of Computing & Informatics (CCI), Drexel University (workshop chair); organizing team members: Florence Hudson, Senior Vice President and Chief Innovation Officer, Internet2; Tim Kraska, Assistant Professor, Department of Computer Science, Brown University; Sam Madden, Professor, Electrical Engineering & Computer Science, Massachusetts Institute of Technology, and Co-Director of the Intel Science and Technology Center (ISTC) in Big Data; and René Bastón, Executive Director, NEBDIH. Logistics and administrative support were overseen by Katie Naum, Program Coordinator, NEBDIH; with assistance from the Metadata Research Center and Drexel University graduate students (Deborah Garwood, Samantha Grabus, Hongwei Liu, Kai Li, and Key Yang), and CCI administrative staff.

Workshop activities took place in the Gerri C. LeBow College of Business at Drexel University, and included 52 participants from a range of sectors and organizations. All four of the NSF/Big Data Innovations Hubs (Northeast, Southeast, Midwest, and West) were represented with active researchers as well as Hub Executive Directors. Among organizations represented were:

- Amazon Web Services, Seattle, WA
- American Museum of Natural History, Drexel University, Philadelphia, PA
- Azavea, Philadelphia, PA
- Baker Botts LLP, Houston, TX
- Bentley University, Waltham, MA
- Boston University, Boston, MA
- Brown University, Providence, RI
- City of New York, NY
- Columbia University, New York, NY
- Comcast, Philadelphia, PA
- DARPA, Arlington, VA
- Data Intensive Cyber Infrastructure/Renaissance Computing Institute (DICE/RENCI), Chapel Hill, NC
- DataScience, Culver City, CA
- Drexel University, Philadelphia, PA
- Experian, Costa Mesa, CA
- GE (General Electric) Global Research, Niskayuna, NY
- Icahn Institute at Mount Sinai, New York, NY
- IDA/Science and Technology Policy Institute, Washington, DC
- IEEE, Piscataway, NJ
- IMS Health, Collegeville, PA
- Independence Blue Cross, Philadelphia, PA
- Internet2, New York, NY
- Lankenau Institute for Medical Research, Wynnewood, PA
- Marinda Management LLC, Carson City, NV
- Massachusetts Green High Performance Computing Center, Holyoke, MA
- MIT, Cambridge, MA
- National Institutes of Health, Bethesda, MD
- Northeast Big Data Innovation Hub, Columbia University, New York, NY
- National Center for Supercomputing Applications, UIUC, Urbana-Champaign, IL
- National Science Foundation, Arlington, VA
- Oracle, Redwood City, CA
- Pennsylvania State University, University Park, PA
- Pfizer, Collegeville, PA
- Rutgers University, State University of New Jersey, New Brunswick, NJ
- South Big Data Innovation Hub
- Tandigm Health, Conshohocken, PA
- Transportation Research Board, Washington, D.C.
- University of Chicago, Chicago, IL
- University of Washington, Seattle, WA
- West Big Data Innovation Hub

A list of workshop participants and their organizational affiliations is included in the appendix.

Workshop Agenda and Content

The workshop began with an introduction and presentation of ground rules for guiding the two days. Jane Greenberg (Drexel University) welcomed participants to Philadelphia and Drexel University, and emphasized the aim for a highly engaged, active, and participatory workshop. Yi Deng, Isaac L. Auerbach Professor and Dean, College of Computing and Informatics at Drexel University, provided a second Drexel welcome and spoke directly to data sharing challenges between industry and academia, drawing from real-life scenarios. Next, René Bastón (NEBDIH) introduced the Northeast Big Data Innovation Hub and the National Science Foundation Big Data Innovation Hubs program in general, followed by Sam Madden (MIT), who introduced the goals of the associated Spokes award. Tim Kraska (Brown University) closed the introductory session by sharing the two-day agenda and reiterating workshop goals.

The workshop design was participant-driven, seeded by guest speaker segments, followed by active breakout group discussions and report-back sessions. Google Docs provided a platform for participants to share their breakout session results. These sessions allowed participants to discuss current challenges hindering data sharing among industry and academia, as well as other organizations, and brainstorm possible solutions and next steps to address these challenges. Brainstorming activity focused on educative, policy, and technical solutions. The themes of standardized legal agreements and policy compliance issues intersected with every discussion.

Presentations over the course of the two-day workshop were as follows:

Day 1: September 29, 2016

- **Robert Cheetham (**Azavea)
  History of Open Data in Philadelphia: Successes, Failures and Lessons Learned
- **Jason Bobe** (Icahn Institute at Mount Sinai)
  What if Your Biology Holds the Key that Protects Others from Disease? Changing the Discourse around Sharing Health Data
- **Joe Chaya** (IMS Health)
  Data sharing challenges and IMS Health
- **John Lee** (Osage Partners)
  Data Sharing
- **Anita Eisenstadt** (Science and Technology Policy Institute)
  Legal Mechanisms to Promote Data Sharing

- **Paul Ragusa** (Baker Botts)
  Data Licensing
- **Greg Madden** (Penn State)
  Towards Seamless Data Sharing
- **Sam Madden** (CSAIL, MIT)
  Towards a technology supported licensing Model
- **Amen Ra Mashariki** (City of New York)
  NYC Analytics

Day 2, September 30, 2016

- **John Brzozowski** (Comcast)
  Data analysis projects: gotchas, remedies, and wish list
- **Yana Kane-Esrig** (Comcast)
  Data analysis projects: gotchas, remedies, and wish list
- **Jane Greenberg** (Metadata Research Center)
  DRYAD
- **Kareem Aggour** (GE)
  Data Sharing Use Case
- **Tony Orsini** (Experian)
  Experian's Predictive Analytics

Main themes covered in the presentations:

- **Data sharing war stories:** Speakers discussed the social impact and research potential of open data alongside funding, time, and human resource requirements associated with sharing sensitive, not fully-open data. Issues of patient identifiability and the concept of informed consent are imperative with medical data, and the competitive nature of data for business start-ups is at odds with the open data priorities of academia.
- **Legal matters, Intellectual property, and policies:** Some aspects of data sharing have advanced via a variety of rights licenses, data access, and reuse policies. For example, data sharing is achievable through various mechanisms, such as standard waivers of copyright, a set of Creative Commons licenses, and non-governmental agreements and policies. Speakers encouraged proactivity in establishing who brings what to the arrangement, policies and procedures, as well as expectations of the ongoing rights and obligations of each party.
- **Infrastructure challenges and opportunities:** Technology-related concerns involved with data sharing include logistics and data transfer speeds, as well as data volume and cost. Issues of institutional trust and long-term compliance with data sharing policies could be addressed through the use of shared repositories.

Additional solutions include mapping between technological solutions and standardized legal agreements to create a data platform that allows researchers to build acceptable licenses from a set of common options, and hosting of data on a server that can enforce these license provisions.

- **Industrial requirements:** Speakers discussed the importance of data format and mechanisms for transferring data, as well as "data hygiene" and options for standardizing certain common data (e.g. addresses). Questions were posed about who hosts data in an industry/academic partnership, and whether raw data or results can be reused for other efforts.
- **City perspective:** City perspectives were presented for both Philadelphia and New York. Opendataphilly.org, spearheaded by Azavea, an independent software innovation company, led to a widespread collective effort toward data sharing, with certain city partners asking to come on board. NYC's Databridge is a cloud-based data repository that facilitates public utility data sharing for the purpose of integration for research, analysis, and operational initiatives. Speakers from both initiatives indicated that the negative consequences of *not* sharing can serve as a powerful incentive for sharing public sector data.

The workshop facilitators organized the first group brainstorming session to generate a list of key data-sharing challenges and consider how the NEBDIH Data Sharing Working Group can help address these challenges. A report-back session followed, and members from each group shared results. Key challenges identified as a result of this activity follow:

- There is tremendous concern around sharing private and competitive information.
- There is a lack of incentives. For many, there is still the absence of a clear rationale for data sharing.
- Overhead expense for the data sharing process is onerously difficult and inhibits success.
- There is concern about liability and preventing mistakes and data misuse. Interconnected here is the limited success with fully tracking data provenance, so that use and misuse cases can be sufficiently accounted for and remedied as necessary.
- There is concern with data as a living entity that has to be updated and, at times, redacted. Here, issues of complying with sharing regulations is a chief concern. A notion of not "throwing it over the wall" was expressed. That is, data sharing shouldn't just be a one-time exchange. Rather, there should be a back-and-forth process, where the recipient of the data gives back cleaned/augmented data to the data owner, and the owner continues to deliver updates and additions to the data over time.

- Data sharing is fraught with complex regulations governing data use.

These initial topics seeded a second breakout session, allowing groups to have a more reflective discussion and further consider where the NEBDIH Data Sharing Working Group, together with researchers engaged with the newly awarded Spoke project, can target next steps. The report-back sessions for this session focused on: **1)** incentives, **2)** overhead costs and making data-sharing more expedient, and **3)** complex regulations governing the use of data. Recommendations falling into these three areas follow:

- **Incentives:** Group ideas included monetary incentives, reducing duplication of effort, avoiding bad PR, and using the Big Data Hubs as an incubator for incentives. Innovation and a common goal can also serve as powerful incentives, with innovation leading researchers to do something new with shared data, and a data commons facilitating progress towards, for example, cancer research. An important key to incentivizing data sharing is to involve people with similar mindsets, who can show the value of data sharing through incremental data-exchange steps, building trust with other organizations over time. To move forward, it is important to make it easy to share data, potentially through co-location, such as a data commons or cloud.
- **Overhead costs and making data-sharing more expedient:** Group members suggested a machine-actionable participant consent form, using pre-defined clauses in comprehendible language. A sandbox environment that enables information access and usage protection was also proposed.
- **Complex regulations governing the use of data:** One of the biggest challenges is enforcing state and federal level regulations. Regulations are frequently viewed as guidelines, so group members see the need to develop a mechanism or application to ensure compliance. One step toward encouraging compliance is to review existing data-driven licenses in order to develop more generalized models.

Summary Observations

As part of the workshop wrap-up, participants reflected on prevalent themes and shared their observations. Cross-cutting themes observed over the course of the two-days, via both presentations and discussions, were:

- **Data sharing motivations differ among stakeholders, although resource concerns are cross-cutting.** The workshop activity made clear that, although the motivation for data sharing differs among stakeholders, there is tremendous concern across all constituents about financial and human resources requirements necessary for ensuring a successful data sharing transaction.

- **Potential gains outweigh the risks.** Another telling observation was that PII and privacy issues clearly present legal and policy oriented data sharing challenges, although the solutions sought are technical. There was substantial agreement that wins, such as curing a disease, or preventing a disaster, far outweigh current risks (e.g., having a person's identity shared). Related here is that the more frequently and broadly high-impact data sharing can be demonstrated, the easier it will likely be to justify resource allocation toward developing tools and protocols for streamlining data sharing.
- **An Integrated policy/technology framework is needed to expedite the data sharing process.** Another thematic observation focused on the need to integrate policy and technological solutions—a topic that can inform a future workshop. An integrated framework will speed up the data sharing process. This approach could drastically reduce the time currently required to procure a successful agreement, and help avoid cases where, by the time the data agreement is secure (e.g., six months or more), the initial value has drastically diminished for all parties involved.

As the workshop closed, participants collectively brainstormed system requirements for the integrated framework, and identified initial action items that both workshop participants and Spokes grant researchers can collaboratively pursue. The system requirements are listed directly below.

Suggested requirements for a prototype system

*Note: This list is the result of a brainstorming session. The following list includes suggestions from the participants, and is not intended to be comprehensive.

A system facilitating data sharing among academia, industry, and/or other constituents should:

- Be user friendly and have language common to all constituents.
- Have templates enabling data sharing as the default.
- Allow data creators and collectors to "opt-out" of an agreement for data sharing, rather than opt in; and requires data creators and custodians to provide a statement of why data cannot be shared.
- Offer different flavors or degrees of opting out, such as embargo periods, or allow selected files or segments of a file to be shared, or not.
- Include an interoperable metadata framework and use standardized ontologies that lend toward data sharing.
- Expediently determine risk assessment against the benefits of data sharing. The system might generate a risk score. The DataTags project at Harvard has developed a confidence-level classification systems for sharing sensitive data, and may have implications for determining risk (Sweeney, et al, 2015).

Action Items

Potential action items cover the areas of: **1)** data gathering and inquiry, **2)** community building, and **3)** educational resource development and outreach. The following action items are initial steps that the workshop participants and Spokes grant researchers may collaboratively pursue, with the NEBDIH data sharing working group helping to facilitate communication among these groups, and also with the other NSF Big Data Innovation Hubs.

Data gathering and inquiry

- Gather documentation from the cases presented in the workshop for further examination.
- Gather information on additional data sharing cases that report challenges, identify obstacles, demonstrate successes, and lay out data sharing solutions.
- Gather examples of licenses and agreements, and look for common clauses in data licensing language
- Examine successful data-sharing repositories. The Re3data.org repository lists many repositories, from which a sample may be drawn.
- Examine the Creative Commons model more thoroughly for developing different licenses.

Educational resources and outreach

- Develop a set of principles that can be used to develop a positive perspective on data sharing and motivate successful partnerships.
- Develop a Data Licensing 101 course that can be accessible to all constituents.
- Reach out to institutions to offer Data Sharing 101 or similar educational offerings. Discussion of this idea included the idea that considered that universities and other research centers may consider Data Sharing 101 as part of an onboarding requirement. It was also suggested that universities might institute a certification program, similar to IRB (internal review board) requirements, that faculty and students pursuing research must complete.
- Link data sharing information to organizational and university mission statements as impetus to facilitate data sharing “for the greater good.”

Community building

- Develop a stakeholder roster of: **1)** people who are willing and able to disclose data sharing lessons learned (successes and failures), and **2)** the larger community of constituents facing challenges and seeking to improve the current status quo. The roster could allow people to register themselves, and also serve to match-make people, ideas, and resources.

- Develop an information-sharing community that is a safe space for people coming from government, academia, industry, and nonprofits, in order to share case studies (and not wind up on Gawker). National Federation of Advanced Information Services (NFAIS) provides one example, as a participant-driven organization with limited sponsorship.
- Engage and work with large communities seeking to share data. Larger groups will help gain momentum and enable greater impact via win/win cases (example communities include Smart Grid and Precision Agriculture Projects).
- Survey members across any organization to "start the conversation" and develop a list of stakeholders in one's organization.
- Support and communicate with funding agencies that are behind data sharing policies.

The workshop closed with participants gathering for a group photograph.

The photo and workshop can be found: http://nebigdatahub.org/datasharingws/.
Workshop announcement page: http://cci.drexel.edu/mrc/news/2016-08-north-east/.
Workshop agenda and slides: http://cci.drexel.edu/mrc/news/2016-11-bigdatahubworkshop/.

Reference and Additional Resources

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, 154, 72-80.

Metadata Research Center, College of Computing & Informatics, Drexel University: http://cci.drexel.edu/mrc/.

Northeast Big Data Innovation Hub: http://nebigdatahub.org/.

Northeast Big Data Innovation Hub Data Sharing Ring/Working Group: http://nebigdatahub.org/data-sharing/.

National Federation of Advanced Information Services (NFAIS): http://www.nfais.org/.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51-59.

Sweeney, L., Crosas, M., & Bar-Sinai, M. (2015). Sharing Sensitive Data with Confidence: The Datatags System. Technology Science.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., ... & Turetken, O. (2014). The current state of business intelligence in academia: The arrival of big data. Communications of the Association for Information Systems, 34(1), 1.

Appendix: Workshop Participants

Name/Institutional Affiliation

Kareem Aggour/GE Global Research
Larry Alexander/Drexel University
René Bastón/Northeast Big Data Innovation Hub
Carsten Binnig/Brown University
Jason Bobe/Icahn Institute at Mount Sinai
John Jason Brzozowski/Comcast
Joseph Chaya/IMS Health
Robert Cheetham/Azavea
Mike Conway/DICE/RENCI
Melissa Cragin/UIUC/NCSA
Thomas DeChiaro/Drexel University
Yi Deng/Drexel University
Robert Downs/Columbia University
Anita Eisenstadt/IDA/Science and Technology Policy Institute
Steve Elliott/Amazon Web Services
Cheryl Flannery/Pfizer
Dan Ford/Azavea
Gary Gabriel/Oracle
Deborah Garwood/The American Museum of Natural History
John Goodhue/Massachusetts Green High Performance Computing Center
Dave Goodsmith/DataScience
Samantha Grabus/Drexel University
Jane Greenberg/Drexel University
Kathy Grise/IEEE
Vasant Honavar/Penn State
Bill Howe/University of Washington
Florence Hudson/Internet2
Kevin Jorissen/Amazon Web Services
Yana Kane-Esrig/Comcast
Tim Kraska/Brown University
Sharon LaDay/Marinda Management LLC
Jack Ledbetter/Experian – Marketing Services
John Lee/Osage Partners
Meredith Lee/West Big Data Innovation Hub
Kai Li/Drexel University
Xia Lin/Drexel University

Hongwei Liu/Drexel University
Greg Madden/Penn State University
Samuel Madden/MIT
Amen Ra Mashariki/City of New York
James Mielke/Rutgers University
Deb Morley/Drexel University
Katie Naum/Northeast Big Data Innovation Hub
Vivek Navale/National Institutes of Health
Amy Nurnberger/Columbia University
Tony Orsini/Experian
David Oury/Bentley University
Norma Padron/Lankenau Institute for Medical Research
David Plazak/Transportation Research Board
Paul Ragusa/Baker Botts LLP
Renata Rawlings-Goss/South Big Data Innovation Hub
Jennifer Roberts/DARPA
Tim Siftar/Drexel University
Aaron Smith-McLallen/Independence Blue Cross
Julia Stoyanovich/Drexel / CCI
Aleister Saunders/Drexel University
Ian Swanson/DataScience
Antonio Tedesco/Tandigm Health
Vas Vasiliadis/University of Chicago
Ke Yang/Drexel University
Saul Youssef/Boston University
Fen Zhao/NSF CISE/OAD