Integrating Approaches to Privacy across the Research Lifecycle

September 24-25, 2013 — Harvard University

Participant Position Statements

John M. Abowd, Edmund Ezra Day Professor of Economics, Department of Economics, Cornell University

September 20, 2013 — Quoting from the introduction to the Cornell NSF-Census Research Network proposal (I am the lead PI; co-PIs are William Block, Ping Li, and Lars Vilhuber; NSF 1131848, awarded September 19, 2011.)

As the National Science Foundation (NSF) recognized with its January 18, 2011 directive that all proposals must include a detailed, viable data management plan, the acquisition, archiving and curation of scientific data is vital to the integrity of the entire process. The test is not "can the next researcher reproduce current results," rather it is "can a researcher working 50 or 100 years from now recover and correctly re-use the original data." Libraries have performed the curation (or preservation) function for millennia. Social scientists recognized the importance of data management decades ago when the Inter-university Consortium for Political and Social Research (ICPSR) was formed, and again a few decades later when NSF funded major social science data initiatives like Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota and the Research Data Centers (RDC) at the Census Bureau.

The Foundation's efforts have been very successful. Figure 1 shows the overall distribution of data sets used in current and historical RDC projects. It summarizes 1,505 project-dataset pairs.[1]1 Fully 71% of all project-datasets use economic micro-data. Such data are primarily the establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD). With the exception of the recently-released Synthetic LBD, there are no public-use micro-data for these establishment-based products. They form the core of the modern industrial organization studies begun by Dunne, Roberts and Samuelson (1989) and Olley and Pakes (1996) as well as

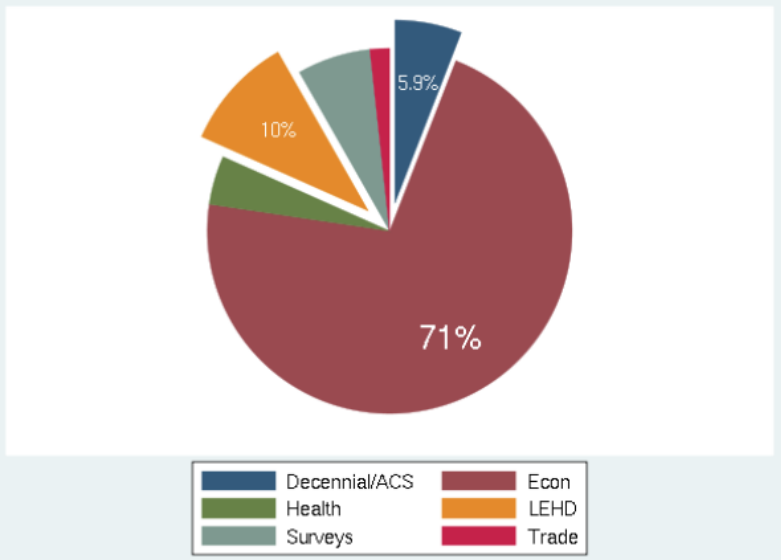

Figure 1 Percentage Use of Confidential Datasets in the RDCs

[1] Many projects use multiple datasets.

modern gross job creation and destruction in macroeconomics (Davis and Haltiwanger and Schuh, 1996; Business Dynamics Statistics[2]). The next most used data come from the Longitudinal Employer-Household Dynamics Program, a longitudinally integrated employer-employee database that was created following a joint Census Bureau-NSF investment in 1999 (Abowd *et al.* 2009). Somewhat surprisingly, only about 6% of the project-dataset pairs involve confidential Decennial/American Community Survey (ACS) data. Public-use decennial files have existed for decades, but we predict that confidential data use will grow substantially now that geographically detailed ACS files are part of the RDC-accessible micro-data.

Over the course of the last decade a framework for providing access to the confidential micro-data that form the basis for the Census Bureau's major data products has emerged. This framework is consistent with the statutory obligations of the Bureau's co-custodians that research use of the micro-data be consistent with the enabling legislation for each constituent data source and that the appropriate administrative review occur prior to the onset of new research. This framework is currently the best available political compromise, but it can be considered neither permanent nor durable. To the extent that the next generations of social scientists build their careers on the basis of original discoveries emanating from these data, a regulatory consensus must emerge that treats the underlying data as a vital scientific asset. When this consensus occurs, it will be too late to begin the curation process.

ICPSR is now the largest social science data repository in the world with over 500,000 data sets in its collection, including a growing inventory of restricted-access datasets.[3] IPUMS and IPUMS-International are the definitive sources for household micro-data originating from population censuses around the world, including projects for which IPUMS-International is the custodian of a foreign nation's confidential micro-data.[4] Researchers working at the Census Bureau and in Census RDCs have acquired and archived a very substantial collection of micro-data that are now used routinely for scientific research in economics, sociology, demographics, environmental science, health, and other fields. But there is a fundamental, and critical, difference between the approach taken by ICPSR and IPUMS as compared to the approach taken by the Census Bureau and other governmental agencies that provide research access to confidential micro-data. The curation function is either absent or woefully neglected. Consequently there is a substantial risk of breach of the scientific integrity of the research process itself because the findings that are reported in the peer-reviewed journals are based on analyses of the confidential restricted-access data, but only public-use data are released for open scrutiny. It is the confidential data themselves that must be curated, not just the disclosure-limited public-use products that this research produces, in order to afford future generations of scientists the same ability to scrutinize this work as many generations have had for work done on the major public-use data products developed in the last 50 years. The statutory custodians of the restricted-access data, in most cases U.S. government agencies, need substantial help from the scientific community in order to ensure that vital research data

[2] See http://www.ces.census.gov/index.php/bds/bds_overview, cited on February 14, 2011.
[3] See http://www.icpsr.umich.edu/icpsrweb/ICPSR/org/index.jsp, cited on February 15, 2011.
[4] See https://international.ipums.org/international/about.shtml, cited on February 15, 2011.

they have now acquired are properly curated.

The problem has been caused by a subtle but pervasive barrier to effective application of current best-practice long-term data management systems. When conventional repositories like ICPSR and IPUMS have attempted to apply the acquisition, archive and curation processes developed for public-use data directly to restricted-access data, the management of restricted-access data adds an additional layer, called stewardship, to the accepted practices. The data archive takes physical custody of a certified-true copy of the confidential data under the terms of a restricted-access data provider agreement with the statutory custodian. This agreement establishes the statutory custodian's legal authority to grant physical data custody to the archive and delineates the terms and conditions of future use, including any disclosure limitation protocols that must be used. At the same time, the archive acquires or creates the metadata that are essential to the curation process. From this point forward, management of the restricted-access data is very similar to management of public-use data. And, in particular, many resources from the data archive and the research community can be used to enhance the curation process.

But if the conventional archive cannot take long-term custody of the original data, this model fails. The Census Bureau and many other government agencies in this country are prohibited by statute from granting an archive like ICPSR or IPUMS long-term physical custody of the data. Both micro-data and metadata are locked up and inaccessible. New confidentiality protection methodologies, developed in part at Cornell under previous NSF awards and Census Bureau contracts to PIs Abowd and Vilhuber, have unlocked large amounts of data for public-use but the structured metadata has not kept pace. Because the Census Bureau retains custody of both the confidential data and critical metadata, a substantially modified curation protocol is required to ensure that the actual inputs to published research are preserved. This protocol is not fully developed. The first goal of our proposal is to design and implement a Comprehensive Census Bureau Metadata Repository (CCBMR) that can be managed outside of the Census Bureau's secure computing domains while maintaining all of the capabilities that a conventional repository would possess. At present, NSF-supported researchers cannot comply fully with the Data Management Plan requirement if they use confidential Census Bureau micro-data. Successful design and execution of the CCBMR would solve that problem.

It is not sufficient to simply build these tools. The next generation of social scientists must be trained to contribute to the repository and use it effectively. Such training naturally supplements instruction that the PIs have already developed on the data production process that drives the Census Bureau's RDCs. The primary engine for that training is a Ph.D. level course (INFO 7470) that was developed with prior NSF support and has been offered every 2 or 3 years to graduate students and faculty around the country. The second goal of our proposal is to design and implement an enhanced course that would be updated annually. At present the community of scholars sufficiently familiar with the Census Bureau's data production processes to successfully mentor new scholars is too small. Successful design and execution of the enhanced INFO 7470 would double or triple that stock over the life of the grant.

Ken Carson, Assistant Provost for Research Policy, Harvard University

Personal Position Statement provided to generate discussion at the Privacy Workshop; not a statement of the Office of the Vice Provost for Research

On September 20, 2013 the NIH released for comment its draft Genomic Data Sharing Policy https://www.federalregister.gov/articles/2013/09/20/2013-22941/draft-nih-genomic-data-sharing-policy-request-for-public-comments#h-23 .

The draft policy gives lip service to the tension between the importance of increasing and accelerating access to genomic research data and the rights of research subjects, but it offers no innovations in managing the tension. Instead, it directs researchers and their institutions to follow the data security policy requirements of the Common Rule, 45 CFR 46, and HIPAA.

The draft policy also acknowledges that for "studies using data or specimens collected *before* the effective date of this Policy there may be considerable variation in the extent to which data sharing and future genomic research was addressed within the informed consent materials for the primary research. In these cases, an assessment by an IRB…is essential to ensure that data submission is not inconsistent with the informed consent provided by the research participant." IRBs are in the business of ensuring that informed consent is obtained; is this a different mandate?

IRBs are also charged with evaluating proposed genomic data submissions and assuring that, among other things,

- Risks to individuals and their families associated with data submitted to NIH-designated repositories were considered;
- To the extent relevant and possible, risks to groups and populations associated with data submitted to NIH-designated data repositories were considered

Are IRBs to consider these risks as they arise from data privacy problems that may be posed even if HIPAA and Common Rule data handling measures are followed?

In summary, researchers and their institutions are being thrown new challenges and responsibilities without adequate support or guidance for meeting them.

Robert Gellman, Privacy and Information Policy Consultant
419 Fifth Street SE, Washington, DC 20003
202-543-7923 | bob@bobgellman.com | www.bobgellman.com

Thoughts, Premises, and Unsupported Assertions

1. Everyone has to follow the law about processing of personal data. But for the most part in the US, there is no law. Everyone has to comply with any contractual agreements about the processing personal data, but contractual obligations may also be rare.

2. Data is an asset. I don't think that needs an explanation.

3. Data is a curse. If you have personal data, you have to comply with the terms under which you obtained the data. You have to take responsibility for the use of the data by your staff. You have to provide reasonable administrative, technical, and physical measures to protect the data. You may have to give data subjects rights of access and correction. You may be subpoenaed to turn over the data for administrative, civil, or criminal matters. You may have obligations (and potentially very expensive obligations) in the event of a security breach. You may have to make decisions about whether and when to share the data for research and other purposes, and you may be liable for those decisions and for other choices about data processing. You may have to account for any disclosures of the data. You have to provide for long-term storage and control over the data.

4. Most researchers processing personal data probably do not have written privacy and security policies. Most IRBs probably do not require (and cannot evaluate) research privacy and security policies. There are no (or inadequate) standards for privacy or security of research data.

5. Privacy laws generally bend over backwards to support research uses of personal data. Yet the research "industry" is only one front-page scandal away from being subject to stringent legislative rules. There are no "industry" standards for privacy or security.

6. One idea that does not solve all problems:

Legislative Proposal for Deidentification of Personal Data. *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010). Deidentification is one method for protecting privacy while permitting other uses of personal information. However, deidentified data is often still capable of being reidentified. The main purpose of this article is to offer a legislative-based contractual solution for the sharing of deidentified personal information while providing protections for privacy. The legislative framework allows a data discloser and a data recipient to enter into a voluntary contract that defines responsibilities and offers remedies to aggrieved individuals. This idea won't solve all problems with the sharing of deidentified data for research or other activities, but it would establish rules and enforceable standards under a statutory scheme. The basic idea is for a federal law, but a state could enact the proposed statute for data within its borders. http://bobgellman.com/rg-docs/RG-Fordham-ID-10.pdf.

Raquel Hill
Associate Professor of Computer Science, Indiana University
Visiting Scholar, Center for Research on Computation and Society, Harvard University

Balancing the Interests in Developing and Sharing Behavioral Science Data

The sharing of medical data has many advantages, including: the creation of a unified data display for clinicians, the development of predictive and diagnostic support systems, reductions in institutional costs, and improvements in medical care. Medical data are often not shared with external parties because even when data is de-identified per HIPAA Safe Harbor rules, sharing of such data may introduce privacy risks. This is because privacy does not only depend on de-identified data but also on context-specific information, such as shared demographic data of subjects, presence of fields that may be linked across other existing databases, social relationships among subjects, and profiles of the data recipient.

Preserving the privacy of medical-related data is even more challenging when the data are obtained from studies that create unique profiles of individual participants. These participant profiles can be so unique that traditional anonymization techniques cannot be used to generalize and de-identify the record. Therefore sharing these data with external parties may become a lengthy process of negotiating specific use agreements. Sharing of the data among researchers within the organization that owns the data also risks privacy. Even when traditional identifiers are removed, the uniqueness of these records may make re-identification easy for anyone who has access to the complete record.

Privacy is important to behavioral science researchers, because they depend upon participants providing accurate answers about their personal behaviors. In prior work [1], we've evaluated two datasets from the Kinsey Institute. These datasets were specifically collected as part of research projects designed to enhance the understanding of behavioral and psychosocial factors related to risk for human immunodeficiency virus (HIV) and other sexually transmitted infections (STI) as well as other risks to well-being. As sexuality-related data are considered sensitive, they require protection of both confidentiality and privacy|-- a shared feature with other types of health data. If participants provide inaccurate information, researchers may make incorrect conclusions with consequences for public health. The protection of data like these serves two purposes: to protect participants' privacy and, through that, to help encourage them to give accurate answers to increase the quality of research. For participants to give accurate information, they must trust that the researchers will strive to protect their information from breaches in confidentiality and invasions of privacy. Researchers must balance three interests in the development and sharing of datasets: 1) the desire to collect data on a range of variables that alone or in combination may risk re-identification; 2) the increasing pressure from the scientific and medical communities as well as funding and regulatory agencies that datasets be made available to others; and 3) human subjects concerns such as potential breaches in confidentiality and privacy.

References
[1] A. C. Solomon, R. Hill, E. Janssen, S. Sanders, J. Heiman, Uniqueness and How it Impacts Privacy in Health-Related Social Science Datasets, In the Proceedings of the *ACM International Health Informatics Symposium (IHI 2012)*, January 28-30, 2012, Miami Florida.

Murat Kantarcioglu, University of Texas at Dallas

Comprehensive Risk Management Framework for Data Privacy

Many human actions ranging from oil extraction to airline travel, involves risks and benefits. In many cases, such as trying to develop a aircraft that may never malfunction, avoiding all risks are either too costly or impossible. At the same time, major risks can be mitigated using cost effective methods (i.e., using seat belt to reduce injury due to car crashes). Similarly, we believe that avoiding all private information disclosure (PID) risks for all individuals would be too costly. Instead, we must carefully balance risks versus benefits in using privacy sensitive data and mitigate various risks through careful use of privacy-preserving technologies (PPT).

To achieve a comprehensive risk management framework, not only we need to estimate PID risks but we also need to understand the potential losses due to privacy violations. Additionally, and more importantly, we need to consider the utility of the data sharing. This implies that the choice of PPT used for mitigating risks while sharing privacy sensitive data must change based on the following factors: 1) The likelihood of PID for each individual due to previously released versions of the dataset and the publicly available data, 2) The severity and/or loss to the individual due to PID. 3) The total expected utility of sharing and using the sanitized data. Since, PID risks depend on the background information of the adversary, various background information scenarios needs to be considered. For example, we may assume that adversary knows the quasi-identifiers for the entire society, e.g. the adversary knows the birthdays and addresses of everyone from the voter registration list.[1] Next, we need to estimate the potential consequences and losses due to a PID. For example, data set that contains sexual preferences and potential re-identification due to release of such data set could create more loss than the release of a data set that contains cholesterol information.

Once all these parameters are estimated, we can reason about the appropriate PPT based mitigation approaches. For example, we may try to set ε in differential privacy techniques based on the risk estimation. In addition, in many cases, simple combination of access control techniques and basic sanitization approaches may be suitable. For example, using secure multi-party based computation techniques, no researcher may see the underlying medical data but instead can get access to the accurate statistical models that they are interested in directly.

Such a risk based decision making also implies that sooner or later some PID will happen. Of course, this means that we need mechanisms to inform and compensate individuals if such PID

event happens. For example, an insurance mechanism may be set up to compensate individuals when PID happens, and the companies and/or organization's using privacy sensitive data may need to pay premiums based on the efficacy of their privacy protection mechanisms.

---

[1] Although differential privacy based techniques can provide some privacy guarantees regardless of the attackers' background information, it may not protect against all types of inference attacks. In addition, it may not be applicable in all contexts.

Henry Lam, Dept. of Mathematics & Statistics, Boston University

As an applied probability and statistics researcher, I came across either simulated or real industrial data in almost every project. I am not a researcher in the privacy area, but I would like to learn about it as I believe it is relevant to any researchers who have to deal with data frequently. I would also like to get some exposure to the research topics in the privacy area.

Wendy Mariner, Dept. of Health Law, Bioethics, and Human Rights, Boston University School of Public Health

One area of my research is the law governing research with human subjects. Another is health information privacy and confidentiality. Of particular interest are controversies over authority to perform secondary and tertiary (and so on) uses (and combinations) of data collected in research studies and in government agency activities (such as health care quality and cost analyses and public health surveillance), and whether and how IT can ensure sufficient security for such uses.

I am intrigued by the NSF Project's title, Privacy Tools for Sharing Research Data, which many law faculty would consider an oxymoron. This particular workshop sounds more promising. I look forward to participating.

Jerry Reiter, Dept. of Statistical Science, Duke University

I thank the conference organizers for giving me the opportunity to offer my thoughts on confidential data dissemination in advance of the workshop. To offer context, I first give some background on my research in this area.

I work extensively on theory and methods for generating synthetic data. The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions. These distributions are specified to reproduce as many of the relationships in the original data as possible. The goal is to protect confidentiality by sacrificing precise information on individual records while preserving global relationships through statistical

modeling.

The U.S. Census Bureau uses synthetic data to disseminate several high profile data products, including the Survey of Income and Program Participation (SIPP), the American Community Survey group quarters (ACSGQ) data, the OnTheMap program and the Longitudinal Business Database (LBD). I advised the Census Bureau on the generation of the SIPP and ACSGQ. I supervised the generation of synthetic data for the LBD, which contains longitudinal data on nearly every business establishment in the U.S. since 1976. The LBD has over 25 million records and dozens of variables. The synthetic LBD is the first-ever public use establishment-level dataset in the United States. It is available for unrestricted download at the Census Bureau website.

I believe that there is enormous value in releasing public use data sets with individual records. Such data enable students to train, which is essential for science to advance. When coupled with query systems or (physical or virtual) secure data enclaves, they can help researchers to be more efficient with their analyses. In particular, data storage and processing of (big) data in enclaves is costly to data stewards, who likely will pass some costs to users. Analysts who have an informed analysis plan based on explorations with synthetic data can improve their efficiency when using the query server or enclave, thereby saving dollars and time.

I am skeptical of the broad usefulness of data dissemination strategies based *only* on differentially private query systems (or any query systems, for that matter). Finite privacy budgets in a query system are, in my opinion, very restrictive for serious data analysts. To understand data, one needs to do many exploratory queries, fit multiple models, examine the validity of those models, etc. I have a hard time imagining that a finite privacy budget can enable all these steps (but certainly I am willing to be proved wrong).

I appreciate the formal guarantees of differential privacy. This line of thinking is a big improvement over the ad hoc risk assessments typically performed by statistical agencies. However, for the type of complex datasets that federal agencies release, I worry that the utility of differentially private synthetic data, at least as generated currently, won't be high enough. Consider the LBD, which includes longitudinal data on payroll and employee size, both highly skewed variables with, in some industries, big outliers. Sensitivity is high in such data (many industry classifications have only a few thousand establishments), so that any noise mechanism presumably results in heavy distortions. Other federal datasets have similar features.

It would be very informative exercise for someone to attempt to create a differentially private synthetic LBD and evaluate the quality of analyses based on it. One could treat the existing public use data, which can be downloaded for free, as real and attempt to synthesize its characteristics. This could go a long way toward convincing agencies of the practical applications of differentially private synthetic data.

I also wonder about the true disclosure risks inherent in releasing a fully synthetic dataset

generated from statistical models, like the synthetic LBD. It would be a fascinating exercise for someone to try to learn the true attributes of a particular establishment from the synthetic LBD. What sort of information does one need to break the protection? What sorts of records are at most risk? What does this imply for future data releases?

If these two projects interest anyone, I'd be delighted to collaborate.

Greta Lee Splansky, Dir. of Operations, Framingham Heart Study, Boston University

1. Considerable thought and effort is required to develop procedures that provide balanced approaches to data use and protection of privacy for research participants. One security tool or procedure may not be appropriate for every study design and every data use.

2. Long term research projects require updating of policies and procedures, security tools, and consent information when significant changes in data use or security exposure occur. Even short-term research projects may need to consider the long-term uses of the resulting data unless there is a plan to destroy the data sets and not to share in the future.

3. Accumulations of amended policies and elaborate measures to work around constraints on data use over time may become unwieldy and burdensome for data use committees, for applicants to data repositories, consortia, IRBs and data managers. New approaches may be needed to permit research to go forward efficiently while affording the best possible protection of participant privacy.

At the Framingham Heart Study (FHS) we have developed approaches throughout 6 decades of operations that may be of use to other studies. For example, we have digitized histories of individual consent options so that data and materials may be distributed in compliance with individual preferences. Researchers may apply for data in an integrated manner that applies to several FHS review committees. FHS has posted its genome-wide association data along with phenotypic data on dbGaP so they can be shared with the scientific community. However, new challenges often arise when scientific proposals entail novel approaches to data use, complex collaborations and consortia, or expansion of study mission.

Peter Suber, Office of Scholary Communication, Harvard University

I come to this workshop as a proponent of open access to research, including research data. At the same time, I believe that medical privacy trumps open access. We should only open up personal medical data with patient consent or when when the data files are sufficiently anonymized. One of the fascinating obstacles here is that the sufficiency of anonymization is both a moving target and a deep problem. In August 2008 that NIH had to take down online two datasets it believed were sufficiently anonymized when it turned out that cleverness and access

to unrelated data allowed the identification of individual subjects <http://www.sciencemag.org/content/322/5898/44.1>. Similar cases have arisen since <http://www.ncbi.nlm.nih.gov/pubmed/22463877>. A pessimist might say that creeping de-anonymization will steadily shrink the scope of open data in fields researching human subjects. An optimist might say that the same cleverness that cooks up new methods of de-anonymization will cook up new methods of anonymization. This may be an arms race with no permanent victor. But I'm interested in the Harvard privacy tools project precisely because it supports optimism.

Adam Tanner, Department of Government Harvard University

My upcoming book, tentatively titled “Behind the Data Curtain,” looks at the business of personal data. It focuses on the real-life impact of personal data gathering by companies. It looks at the human stories of the entrepreneurs and firms gathering personal data and the consumers about whom data is gathered. My research encourages companies to be open about what customer data they aggregate and how they use it. Such openness in turn allows people to choose to what extent they are comfortable in having their data collected.

Salil Vadhan, Vicky Joseph Professor of Computer Science and Applied Mathematics, School of Engineering and Applied Sciences, Harvard University

I am a theoretical computer scientist, and my own research is on differential privacy (along with other areas of theoretical computer science), but here I'd like to raise a broader question for discussion during the workshop: how much "nuance" is feasible to incorporate into the handling of privacy-sensitive research data?

Understandably, there is much attractiveness to simple and unambiguous procedures, such as the stripping of 18 identifiers specified in the HIPAA Safe Harbor. Such rules are easy to follow and enforce. But they are also crude, and can lead to inadequate protections (e.g. high reidentification risk), loss of data utility (e.g. eliminating information that one wishes to study), or both.

I believe that there is no satisfactory, one-size-fits-all set of technical requirements for the sharing of privacy-sensitive data. The standards and solutions that are appropriate for one form of data in one context are typically inapplicable to the others. Solutions should be tailored to the structure of the data (e.g. standard relational microdata vs. social network data vs. text), the sensitivity of the information and potential harms of disclosure, the level of consent obtained from subjects, and the intended recipients of shared data. Indeed, sharing with researchers governed by IRBs, sharing with the public, and sharing under limited data-use agreements should all be treated differently.

How can we take all of this nuance into account in practice? Clearly, a case-by-case review by privacy experts would be far too cumbersome. What sorts of instruments (legal, technological, or other) might be most effective and efficient for this goal? A "data tag" generator, as demonstrated by Merce Crosas on day 1 of the workshop, could be one useful tool. When commenting on the proposed revision of the Common Rule (http://privacytools.seas.harvard.edu/files/privacytools/files/commonruleanprm.pdf), we proposed the creation of a "safe-harbor catalogue" that enumerates acceptable data-sharing mechanisms according to the various dimensions enumerated above, but this raises additional questions regarding the maintenance of the catalogue and whether a "safe harbor" is the appropriate legal instrument. I hope that the workshop discussions will provide a better sense of the range of approaches available for incorporating nuance into the handling of privacy-sensitive data.