Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier

Forthcoming in Berkeley Technology Law Journal 33:2 (2018)

Based on the Tenth Annual Berkeley Law Privacy Lecture
Berkeley Center for Law and Technology
November 16, 2017

Christine L. Borgman, UCLA
Distinguished Professor and Presidential Chair in Information Studies
University of California, Los Angeles
http://christineborgman.info

Abstract

As universities recognize the inherent value in the data they collect and hold, they encounter unforeseen challenges in stewarding those data in ways that balance accountability, transparency, and protection of privacy, academic freedom, and intellectual property. Two parallel developments in academic data collection are converging: (1) open access requirements, whereby researchers must provide access to their data as a condition of obtaining grant funding or publishing results in journals; and (2) the vast accumulation of "grey data" about individuals in their daily activities of research, teaching, learning, services, and administration. The boundaries between research and grey data are blurring, making it more difficult to assess the risks and responsibilities associated with any data collection. Many sets of data, both research and grey, fall outside privacy regulations such as HIPAA, FERPA, and PII. Universities are exploiting these data for research, learning analytics, faculty evaluation, strategic decisions, and other sensitive matters. Commercial entities are besieging universities with requests for access to data or for partnerships to mine them. The privacy frontier facing research universities spans open access practices, uses and misuses of data, public records requests, cyber risk, and curating data for privacy protection. This paper explores the competing values inherent in data stewardship and makes recommendations for practice, drawing on the pioneering work of the University of California in privacy and information security, data governance, and cyber risk.

The world's most valuable resource is no longer oil, but data[1].

If you can't protect it, don't collect it[2].

Framing the problem

Universities are stewards of vast amounts of data. These data provide many new opportunities for research, teaching, administration, partnerships, and strategic planning. Data take many forms, have many origins, and have many uses. Data ownership is rarely clear, especially for research data, and the costs and mechanisms for stewardship are poorly understood. While data are difficult to manage and to govern in any institution, universities face a particularly complex set of responsibilities and risks.

Stewardship of data and of public trust are sometimes asymmetrical. The university community, which includes students, faculty, staff, and many other stakeholders, expects a reasonable degree of confidentiality in their dealings with an institution of research and learning. They also expect the university to respect their privacy and to keep their data secure. Furthermore, faculty and students expect their universities to respect their academic and intellectual freedom while managing and governing data. The public, which extends beyond the university community, expects universities to be fair, transparent, and accountable for resources. Good stewardship means releasing some kinds of data and preventing the release of other kinds of data. The same data may fall into either category, depending on the time, purpose, or entity requesting access. Few universities – or other institutions – have adequate governance mechanisms to address these stewardship challenges effectively.

This broad set of concerns was framed succinctly by the University of California Privacy and Information Security Initiative (PISI) which was charged in 2010 by then-President Mark Yudof to make recommendations for an overarching privacy framework to address the university's statutory and regulatory obligations; governance, implementation, and accountability structures; and policy vehicles for university policy and practice in privacy and information security.[3] As evidence of the importance placed on this effort, the PISI Steering Committee was drawn from the upper echelons of the University, including the Provost, General Counsel, Chief Compliance

---

[1] The world's most valuable resource is no longer oil, but data, THE ECONOMIST, 2017, http://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource (last visited May 16, 2017).

[2] Privacy and security aphorism.

[3] UC PRIVACY AND INFORMATION SECURITY INITIATIVE STEERING COMMITTEE REPORT TO THE PRESIDENT, 1–43 27–28 (2013), http://ucop.edu/privacy-initiative/uc-privacy-and-information-security-steering-committee-final-report.pdf.

and Audit Officer, Chief Information Officer, and representatives from the campuses and the Academic Senate.

In considering its charge, "the Steering Committee was guided by the following principles":[4]

- We must maximally enable the mission of the University by supporting the values of academic and intellectual freedom.
- We must be good stewards of the information entrusted to the University.
- We must ensure that the University has access to information resources for legitimate business purposes.
- We must have a University community with clear expectations of privacy—both privileges and obligations of individuals and of the institution.
- We must make decisions within an institutional context.
- We must acknowledge the distributed nature of information stewardship at UC, where responsibility for privacy and information security resides at every level.

These principles have proved to be robust in the several years since the final report was submitted to the President and the Regents. Most of the recommendations have been adopted and implemented, including appointing Chief Privacy Officers and establishing joint Academic Senate-Administration boards on privacy and information security on each of the ten campuses. At the UC-wide level, the Academic Senate monitors PISI implementation via the UC Academic Computing and Communications Committee (UCACC). Individual campuses have extended the PISI principles in various ways. UCLA, which established a joint Senate-Administration Privacy and Data Protection Board in 2005, extended the PISI principles and recommendations in the Data Governance Task Force Report.[5]

Identifying problems and principles is an essential starting point to address challenges of the day. Applying these principles to solve these problems is much harder. Over the last several years, the complexity of these challenges has become ever more apparent. This paper explores the current landscape of opportunities, responsibilities, risks, and frontiers facing universities in a data-rich world. It draws on the pioneering work of the University of California, one of the world's premier public research universities, at the forefront of both data governance and data exploitation. It also draws on a large body of work on policy and practice for governing access to research data.

The epigraphs on page 1 frame the arguments herein. Data have become the "new oil," as the modern world's most valued commodities. Market leaders, whether in commerce or in higher education, may be those most adept at exploiting data in their realms. As non-consumptive goods, arguably more valuable than the finite supply of oil, data can be mined, combined, and reused for multiple applications over long periods of time.[6] The aphorism, "if you can't protect

---

[4] *Id.* at 8.
[5] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, 1–41 (2016).
[6] CHRISTINE L. BORGMAN, BIG DATA, LITTLE DATA, NO DATA: SCHOLARSHIP IN THE NETWORKED WORLD (2015); CHARLOTTE HESS & ELINOR OSTROM, UNDERSTANDING KNOWLEDGE AS A COMMONS: FROM THEORY TO PRACTICE (2007).

it, don't collect it," has circulated in the privacy, security, and hacker communities for a decade or more. Leaking data can be at least as dangerous as leaking oil. For universities to sustain the public trust, and to live by the principles that guided the UC Privacy and Security Initiative, they must address the converse of that aphorism: "if you collect it, you must protect it."

The Data-Rich World of Research Universities

Universities are stewards of many kinds of data, some of which they collect, others that they acquire, and yet others that are byproducts of regular activities. The value of these data, the possibilities for exploitation, responsibilities for stewardship, and the types of associated risks vary immensely.

Intentional data collection is the more obvious sort, such as materials gathered by investigators as part of research projects, and information about current and prospective students gathered by the registrar. These data tend to be governed by established mechanisms such as grant contracts, Institutional Review Boards, HIPAA, and FERPA.[7] At the other extreme is incidental data collection that is difficult to identify or govern, such as that gathered by students, by staff in administrative roles, and by technology such as security cameras controlled within offices or departments. A growing source of incidental data collection is software packages that individuals install on university networks for otherwise legitimate purposes in teaching and research. In between is a vast array of data collection that may be more or less intentional, more or less governed, and whose applications may be more or less foreseeable at the time of collection. These include learning management systems, personnel systems that include faculty dossiers for academic evaluation and promotion, identity cards that encode various privileges (library usage, food service, building access, debit charges, etc.), and much more.

In all of these arenas, data volumes and variety are growing at rates far greater than most administrators or faculty are aware. Those individuals who recognize the value and opportunities in these data are not necessarily obligated to seek permission to exploit them. Third parties outside the university may be the first to recognize data opportunities, and approach individuals at any level of the university for partnerships. Governance mechanisms to assure protection of privacy, academic freedom, intellectual property, information security, and compliance with regulations in the uses of such data are nascent, at best.

Of this immense landscape of data issues in universities, this paper focuses on two exemplars. The first is research data, spanning all academic domains from the sciences, technology, and medicine to the social sciences, humanities, and the arts. While the data management issues in these areas are critical and far from solved, at least two decades of practice and policy inform current discussions. The second is data collected by universities about members of its community, including students, faculty, staff, visitors, patients, and other stakeholders. Data collection about individual persons, both intentional and incidental, is accelerating rapidly with

---

[7] HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996 (HIPPA), P.L. No. 104-191, 110 Stat. 1936 (1996); FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT OF 1974 (FERPA), 20 USCS § 1232g (1974); PROTECTION OF HUMAN SUBJECTS, 45 C.F.R §46 (2009).

the implementation of systems that can exploit "data exhaust" from the activities of individuals.[8] Despite several decades of research on principles and practice for "privacy by design," developers too often default to collecting as much data as possible.[9]

Data that universities collect about their communities is also a large and diverse category. The primary exemplar discussed herein is teaching and student learning, itself an area of data explosion. Fully online courses can capture data on every keystroke of every participant, if they choose to do so, creating rich profiles on individual students and interactions between them. Less obvious is the amount of data produced in hybrid courses, where learning management systems complement interactions in campus classrooms. Students acquire their readings and assignments online, participate in online discussions and other activities, and submit their assignments through these systems, all of which is discretely time-stamped. When learning data are aggregated with other kinds of data that universities hold on their students, extensive profiles result. These datasets can be deployed for learning analytics, institutional reports to government and accreditation agencies, academic research, or for surveillance of activities and behavior.

Research data

Scholars collected research data long before the advent of the scholarly journal, which is barely 350 years old.[10] Data are reported in publications, usually in selected and synthesized forms. Some data are kept for reuse by investigators; other data may become barter in exchange for other data or as invitations to collaborate.[11] Until recently, data were considered part of the

---

[8] VIKTOR MAYER-SCHONBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013).

[9] Victoria Bellotti & Abigail Sellen, *Design for Privacy in Ubiquitous Computing Environments*, *in* PROCEEDINGS OF THE THIRD EUROPEAN CONFERENCE ON COMPUTER-SUPPORTED COOPERATIVE WORK 13–17 SEPTEMBER 1993, MILAN, ITALY ECSCW '93 77–92 (Giorgio de Michelis, Carla Simone, & Kjeld Schmidt eds., 1993), https://doi.org/10.1007/978-94-011-2094-4_6 (last visited Oct 12, 2017); Herbert Burkert, *Privacy-enhancing technologies: typology, critique, vision*, *in* TECHNOLOGY AND PRIVACY 125–142 (Philip E. Agre & Marc Rotenberg eds., 1997), http://dl.acm.org/citation.cfm?id=275283.275288 (last visited Oct 12, 2017); I. Goldberg, D. Wagner & E. Brewer, *Privacy-enhancing technologies for the Internet*, *in* PROCEEDINGS IEEE COMPCON 97. DIGEST OF PAPERS 103–109 (1997), https://doi.org/10.1109/CMPCON.1997.584680; Katie Shilton, *Participatory Personal Data: An Emerging Research Challenge for the Information Sciences*, 63 J. AM. SOC. INF. SCI. TECHNOL. 1905–1915 (2012), http://dx.doi.org/10.1002/asi.22655 (last visited May 17, 2015); Katie Shilton, *Values Levers: Building Ethics into Design*, 38 SCIENCE, TECHNOLOGY, & HUMAN VALUES 374–397 (2013), http://dx.doi.org/10.1177/0162243912436985 (last visited Jan 20, 2017).

[10] The Royal Society, 350 YEARS OF SCIENTIFIC PUBLISHING (2015), https://royalsociety.org/journals/publishing-activities/publishing350/ (last visited Oct 9, 2017).

[11] Stephen Hilgartner & Sherry I. Brandt-Rauf, *Data access, ownership and control: Toward empirical studies of access practices*, 15 KNOWLEDGE 355–372 (1994), https://doi.org/10.1177/107554709401500401.

research process, rather than products to be disseminated. Data release has become a condition of obtaining grants and publishing papers in many domains, especially in the biosciences and medicine. Survey research in the social sciences has a long history of data sharing. In the humanities, archaeology is a growth area for data sharing and archiving.

When datasets were small and locally controlled, issues of data stewardship and governance rarely arose. As datasets became larger, distributed collaborations became more common, and data were born digital, tools to mine and combine data became more sophisticated. These opportunities vary immensely between domains, universities, countries, and cultures, as do applicable policies.[12] As the volume of publicly available research data expands, concerns for stewardship of these data become more urgent.[13]

Scope and Definitions

One part of the challenge in managing research data is the difficulty of defining "research" or "data" succinctly. Information, documents, and materials exist in many forms, and in many states, only a portion of which might be considered research data for the purposes of governance. The Oxford English Dictionary is a good starting place to define concepts such as *research*:

> The act of searching carefully for or pursuing a specified thing or person; an instance of this. Systematic investigation or inquiry aimed at contributing to knowledge of a theory, topic, etc., by careful consideration, observation, or study of a subject. In later use also: original critical or scientific investigation carried out under the auspices of an academic or other institution. Investigation undertaken in order to obtain material for a book, article, thesis, etc.; an instance of this.

Locating a singular definition of *research* used within the University of California proved elusive. At UCLA, for example, the Office of Research Administration lists responsibilities and resources on its website, but does not define *research* in its glossary of terms. Research, like beauty, is often in the eye of the beholder, who may be a grant-funding program manager or an academic personnel officer.[14] One area in which firm definitions are needed are studies involving human subjects. In the U.S., such studies fall under the regulation of the federal Department of Health and Human Services:

> DHHS regulations define *research* as a *systematic investigation*, including research development, testing and evaluation, designed to *develop or contribute to generalizable knowledge* (45 CFR 46.102(d)).

---

[12] BORGMAN, *supra* note 6.

[13] Francine Berman & Vinton G. Cerf, *Who Will Pay for Public Access to Research Data?*, 341 SCIENCE 616–617 (2013), dx.doi.org/10.1126/science.1241625 (last visited Aug 23, 2013); Tony Hey & Anne Trefethen, *Cyberinfrastructure for e-Science*, 308 SCIENCE 818–821 (2005), dx.doi.org/10.1126/science.1110410; Jeremy York, Myron Gutmann & Francine Berman, *What Do We Know About The Stewardship Gap?*, UNIVERSITY OF MICHIGAN LIBRARY (2016), http://hdl.handle.net/2027.42/122726 (last visited Jul 27, 2016).

[14] DONALD STOKES, PASTEUR'S QUADRANT: BASIC SCIENCE AND TECHNOLOGICAL INNOVATION (1997).

If a study meets these requirements and is deemed to involve human subjects, then the protocol must be submitted to the Institutional Review Board (IRB) of the university. Whether a study is considered research or involves human subjects is not always obvious. A systematic study that involves a survey of students for the purposes of university strategic planning is usually not considered research because it is not intended for publication, and thus not for generalizable knowledge. Systematic investigations of human activity that are intended for publication, but that do not require direct contact with individual living persons may or may not be deemed research for the purpose of IRB review. Problems arise when data collected for administrative purposes later are deemed worthy of publication, which is not an uncommon occurrence.

Research data is similarly problematic to define, and is often left undefined in guidelines for releasing or depositing data from a research project. At best, data may be defined by example, such as observations, facts, samples, or records. The definition developed elsewhere is the basis for this discussion: "data refers to entities used as evidence of phenomena for the purposes of research or scholarship".[15] This phenomenological definition covers data in any academic discipline, recognizing that one scholar's signal is another's noise.

Open Access to Research Data

Practices and policies for open access to research data are intertwined with those for open access to scholarly publications such as journal articles. Since the early days of "electronic publishing" in the 1990s, activists have called for open access to scholarly publications as a means to democratize access to information.[16] Open access has taken many forms, such as disseminating pre-prints prior to publication, post-prints after publication, or publishing in journals that are free to read online.[17] A growing number of scholarly books also are being published in open access formats, often with print-on-demand options. Open access increases the dissemination of research, which tends to enhance the visibility of authors and their institutions, so the payoffs are several. Economic models for open access dissemination vary widely, as do stewardship models. Responsibility for access and for sustainability often fall to different parties.[18]

---

[15] BORGMAN, *supra* note 6 at 29.

[16] Stevan Harnad, *The post-Gutenberg galaxy: How to get there from here*, 11 THE INFORMATION SOCIETY 285–292 (1995), https://doi.org/10.1080/01972243.1995.9960203; Stevan Harnad, *Post-Gutenberg galaxy: The fourth revolution in the means of production of knowledge*, 2 PUBLIC-ACCESS COMPUTER SYSTEMS REVIEW 39–53 (1991), http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html.

[17] PETER SUBER, OPEN ACCESS (2012).

[18] Isabel Bernal, *Open Access and the Changing Landscape of Research Impact Indicators: New Roles for Repositories*, 1 PUBLICATIONS 56–77 (2013), dx.doi.org/10.3390/publications1020056 (last visited Jul 16, 2014); CHRISTINE L. BORGMAN, SCHOLARSHIP IN THE DIGITAL AGE: INFORMATION, INFRASTRUCTURE, AND THE INTERNET (2007); Harvard University, OPEN ACCESS POLICIES (2010), https://osc.hul.harvard.edu/policies (last visited Jan 10, 2014); Jennifer Howard, *Open Access Gains Major Support in U. of California's Systemwide Move*, THE CHRONICLE OF HIGHER EDUCATION, 2013; Office of Scholarly Communication, UC OPEN ACCESS POLICY (2014), http://osc.universityofcalifornia.edu/openaccesspolicy/ (last visited Aug

Providing access to research data is often a condition for publishing an article, whether or not the article itself is published in an open access form. Thus, data release usually occurs at the time of submitting a paper for publication. Datasets can be contributed to archives or repositories, which assign them a unique identification number, and that ID is linked to the paper. Ideally, it becomes possible to search for data and identify associated publications, or to search for publications and identify associated datasets.[19] Publishing articles in open access venues and disseminating preprints are more established practices than is open access to data. Data release varies widely by domain, with greatest acceptance in the biosciences and medicine, and by type of data, research method, funding source, and other factors.[20]

A legacy of open access publishing that contributes to stewardship challenges is that the notion of "publication" has become more diffuse. Whether something can be considered a formal publication matters for evaluating scholarship, and thus for hiring, tenure, grant proposals, library collections, and much more. In the print world, publications were more readily distinguishable from "grey literature." The latter category consists of documents such as working papers, reports, pamphlets, and preprints that have scholarly value, but that have not been vetted by peer review or disseminated through a formal publication process. In the online world, versions of scholarly documents proliferate. The same or similar content, often with the same or similar titles and authors, may appear as pre-prints, post-prints, working papers, slide decks, and as the formal "official" version of a publication. Initial versions of documents may or may not become formal publications at a later time. Others may diverge into multiple publications. Choosing which version to cite is a judgment call by the citing author.

The publication versioning problem intersects with the data stewardship problem in at least two ways. One is determining the relationship between a dataset and a publication or other document that describes the dataset. Research projects can generate many versions of publications and many versions of datasets, resulting in a complex array of many-to-many relationships between datasets and publications explaining the context in which they were created. The other problem is conflating data release with "data publishing," which has become popular terminology. Datasets can be assigned Digital Object Identifiers (DOI) when contributed to a data archive or repository, giving them an equivalent technical status for retrieval purposes.[21] However,

---

29, 2013); Randall Munroe, *The Rise of Open Access*, 342 SCIENCE 58–59 (2013), dx.doi.org/10.1126/science.342.6154.58 (last visited Oct 4, 2013); Richard Van Noorden, *Europe joins UK open-access bid*, 487 NATURE 285–285 (2012), dx.doi.org/10.1038/487285a (last visited Aug 21, 2013); JOHN WILLINSKY, THE ACCESS PRINCIPLE: THE CASE FOR OPEN ACCESS TO RESEARCH AND SCHOLARSHIP (2006).

[19] BORGMAN, *supra* note 18; Philip E. Bourne, *Will a biological database be different from a biological journal?*, 1 PLOS COMPUTATIONAL BIOLOGY e34 (2005), https://dx.doi.org/10.1371/journal.pcbi.0010034.

[20] BORGMAN, *supra* note 18; BORGMAN, *supra* note 6.

[21] CODATA-ICSTI Task Group on Data Citation Standards and Practices, *Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data*, 12 DATA SCIENCE JOURNAL 1–75 (2013), dx.doi.org/10.2481/dsj.OSOM13-043; FOR ATTRIBUTION—

assigning a DOI to an object, whether a journal article, conference paper, dataset, presentation slide deck, glass slide, or other entity does not make it a publication in the formal sense of peer review, nor does a DOI make an object worthy of long-term stewardship.[22]

Opportunities in Research Data

Democratizing access to knowledge is among the drivers of open access to publications and to research data. The opportunities in these categories differ in important ways, however. Open access to publications expands readership to audiences far beyond the privileged communities that enjoy access to expensive journals and books through their university libraries. Whether read in the form of pre-print, post-print, or published journal article, open access dissemination of scholarly work has created a vast international audience of interested students, researchers, enthusiasts, patients, parents, and other parties. Having the domain knowledge and linguistic ability to exploit these materials is another matter, but providing access is a good start on the equity issues.

Similarly, open access to research data expands scholarly data resources far beyond the investigators who collected and analyzed them. Others can exploit these data, as intact datasets or in combination with other resources, for many purposes. The barriers of requisite domain knowledge and linguistic skills still apply, but the opportunities to exploit data are suggested to be boundless. Among the policy drivers commonly cited are transparency, to allow others to inspect and evaluate findings; reproducibility, to verify findings by repeating a study; and reuse, whether as an independent dataset or aggregated with other data. Accountability to taxpayers, in the case of public funding, is also mentioned frequently.[23]

The promises of open access to research data are vast, although mired in hyperbole. Vast stores of research data are predicted to accumulate through open access policies enforced by publishers, funding agencies, and government directives, and through voluntary participation. These stores can be mined and combined by anyone, at least in principle, leading to new research findings,

---

DEVELOPING DATA ATTRIBUTION AND CITATION PRACTICES AND STANDARDS: SUMMARY OF AN INTERNATIONAL WORKSHOP, (Paul F. Uhlir ed., 2012).

[22] BORGMAN, *supra* note 6; Christine L. Borgman, *Data Citation as a Bibliometric Oxymoron*, *in* THEORIES OF INFORMETRICS AND SCHOLARLY COMMUNICATION 93–115 (Cassidy R. Sugimoto ed., 2016).

[23] Access to scientific information in the digital age, EUROPEAN COMMISSION, RESEARCH, SCIENCE AND SOCIETY (2008), http://ec.europa.eu/research/swafs/index.cfm; BORGMAN, *supra* note 6; Geoffrey Boulton, *Open your minds and share your results*, 486 NATURE 441–441 (2012), dx.doi.org/10.1038/486441a (last visited Aug 20, 2013); GEOFFREY BOULTON ET AL., SCIENCE AS AN OPEN ENTERPRISE (2012), http://royalsociety.org/policy/projects/science-public-enterprise/report/ (last visited May 24, 2013); ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, OECD PRINCIPLES AND GUIDELINES FOR ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING 1–24 (2007), www.oecd.org/dataoecd/9/61/38500813.pdf; John P. Holdren, MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: INCREASING ACCESS TO THE RESULTS OF FEDERALLY FUNDED SCIENTIFIC RESEARCH (2013).

new innovations, new companies, and new market sectors.[24] European policy presentation at a recent Research Data Alliance meeting suggested that "By 2020, the European Data Economy in the most favourable scenario could contribute up to 4% of EU GDP." [25]

In practice, however, considerable investment is required to make research data useful to anyone beyond the original data collectors. Whereas most scholarly documents can be read and understood as independent units, the same is not true of data. A dataset alone, without accompanying documentation of the research methods by which it was created, analysis and interpretation of the findings, and associated context such as instruments, models, and software, may be little more than a string of numbers. The better documented and curated, the more useful any given set of data will be to others.[26]

Grey Data: Academic, Administrative, and Instructional

"Grey data" is proposed as an umbrella term to describe the vast array of data that universities accumulate outside the research realm. Analogous to grey literature, explained above, these are useful data that have not been vetted by peer review, or perhaps by any other governance mechanism of the university. Grey data have become critical to a university's ability to

---

[24] Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, 16 WIRED, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory; Peter Arzberger et al., *An International Framework to Promote Access to Data*, 303 SCIENCE 1777–1778 (2004), dx.doi.org/10.1126/science.1095958; Geoffrey Boulton et al., *Open Data in a Big Data World: An International Accord*, (2015), https://www.icsu.org/publications/open-data-in-a-big-data-world (last visited Oct 21, 2017); Kenneth Cukier, *Data, data everywhere*, THE ECONOMIST, 2010, http://www.economist.com/node/15557443 (last visited Jul 27, 2013); Dawn Field et al., *'Omics Data Sharing*, 326 SCIENCE 234–236 (2009), dx.doi.org/10.1126/science.1180598 (last visited Jun 13, 2013); Brooks Hanson, Andrew Sugden & Bruce Alberts, *Making Data Maximally Available*, 331 SCIENCE 649–649 (2011), dx.doi.org/10.1126/science.1203354; Holdren, *supra* note 23; S. D. Kahn, *On the Future of Genomic Data*, 331 SCIENCE 728–729 (2011), dx.doi.org/10.1126/science.1197891; J. Palfrey & J. Zittrain, *Better Data for a Better Internet*, 334 SCIENCE 1210–1211 (2011), dx.doi.org/10.1126/science.1210737 (last visited Oct 26, 2017); O. J. Reichman, M. B. Jones & M. P. Schildhauer, *Challenges and Opportunities of Open Data in Ecology*, 331 SCIENCE 703–705 (2011), dx.doi.org/10.1126/science.1197962; The Global Alliance for Genomics and Health, *A federated ecosystem for sharing genomic, clinical data*, 352 SCIENCE 1278–1280 (2016), dx.doi.org/10.1126/science.aaf6162 (last visited Oct 27, 2017); Special Issue: Dealing with Data, 331 SCIENCE 692–729 (2011), http://science.sciencemag.org/content/331/6018.

[25] Celina Ramjoué, BUILDING A EUROPEAN DATA ECONOMY: THE ROLE OF RESEARCH DATA. (2017).

[26] BORGMAN, *supra* note 6; Irene V. Pasquetto, Bernadette M. Randles & Christine L. Borgman, *On the Reuse of Scientific Data*, 16 DATA SCIENCE JOURNAL (2017), dx.doi.org/10.5334/dsj-2017-008 (last visited Mar 29, 2017).

"innovate, enhance, and execute our core missions of education, research, and services".[27] Some of these data are collected for mandatory reporting obligations such as enrollments, diversity, budgets, grants, and library collections. Many types of data about individuals are collected for operational and design purposes, whether for instruction, libraries, travel, health, or student services. Universities are increasingly aware of the asset value of data about their communities. Some of these data have legal encumbrances for compliance purposes, but many are collected for reasons of internal management and external competitiveness. Outside entities also see the value in these data, whether through explicit partnerships with universities to exploit data, or by collecting data on users of their products.

The drivers of data collection in universities are many, not the least of which is "market-based solutions" as a response to the lack of funding for public colleges and universities. Higher education reform is being defined in "highly economistic terms" leading to "measurement panic".[28] University administrators may be given statistical benchmark targets for enrollments, time to degree, retention, diversity, and other countable factors, not unlike performance targets in private business. When higher education is viewed more as a job track than as an investment in a democratic citizenry, market-driven measurement may be an inevitable result. Competition looms everywhere.

Collecting Grey Data

Universities always have collected data about their communities, their operations, and their services – as do businesses, governments, and public service sectors. As daily activities of teaching, learning, research, and operations have moved online, the "volume, variety, and velocity" of data collection have exploded. [29] The uses of digital data from online networks differ from data collected offline in at least two respects. One is that discrete data elements become far more valuable when combined with other data. Information gathered about student performance in a single course, once aggregated with data on performance in other courses, test scores, social media activity, library usage, and dietary habits, for example, yield rich profiles on individuals. The other difference between offline and online collection is that many more people have access to those data. In the past, an individual instructor knew little about students enrolled in her course beyond the list provided by the registrar. Now the instructor may be given profiles on each student to track progress. Academic counselors, student advising staff, instructional designers, registrars, department chairs, deans, provosts, and many others may also have access to these data.

The pervasiveness of information technologies has accelerated over the course of several decades, much of which originated in university environments. Today's senior faculty have lived

---

[27] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5 at 3.

[28] Sanford F. Schram, *The Future of Higher Education and American Democracy: Introduction*, 36 NEW POLITICAL SCIENCE 425–437, 427–430 (2014), https://doi.org/10.1080/07393148.2014.954805 (last visited Aug 12, 2017).

[29] Doug Laney, *3D data management: Controlling data volume, velocity and variety*, 6 META GROUP RESEARCH NOTE 70 (2001).

through eras of mainframe computers, minicomputers, desktop personal computers, and ubiquitous mobile devices such as laptops, tablets, and smart phones. They have adapted their research and teaching practices to accommodate, if not to incorporate these technologies. Instrumentation large and small is deeply embedded in the practice of many domains, ranging from space telescopes to sensor networks to nanotech devices. In the early days of portable technologies, instruction practices excluded these devices from the classroom, asking students to leave their calculators, cell phones, and laptops at home, or at least out of sight. While some faculty continue to bar mobile technologies from classrooms, most have embraced tools such as learning management systems (LMS) that support course websites, links to reading materials, discussion groups, and authentication to library and enrollment services. Pedagogy has shifted rapidly over the last decade from rejecting or ignoring students' uses of information technologies to embracing "cyberlearning," both for the analytical data generated and for the ability to adapt instruction to students' behavior.[30]

Opportunities in Grey Data

As cited above, data have become the "new oil" that drives commerce and competition.[31] Google, Amazon, Facebook, and many other companies have built financial empires by collecting and combining personal data. These data are used to profile individuals, segment the population into discrete units, and present information highly selectively. They can also be used to monitor or predict behavior, resulting in closer observation for illicit or suspicious activities, or for auspicious moments to present advertisements, news, or other content. Many decisions are made about people on the basis of their online traces.

Universities, often with commercial partners, are exploiting data about individuals in similar ways. By collecting detailed data on individual student performance, some universities are creating an individualized "learning path" for each student, with various benchmarks toward degree completion.[32] Other institutions are constructing profiles that assign students to one of three categories that predict success, such as the green, yellow, and red "Stoplight" system. Some

---

[30] BEN WILLIAMSON, BIG DATA IN EDUCATION: THE DIGITAL FUTURE OF LEARNING, POLICY AND PRACTICE (1 edition ed. 2017); Goldie Blumenstyk, *As Big-Data Companies Come to Teaching, a Pioneer Issues a Warning*, THE CHRONICLE OF HIGHER EDUCATION, 2016; Paul Voosen, *Big-Data Scientists Face Ethical Challenges After Facebook Study*, THE CHRONICLE OF HIGHER EDUCATION, 2014; CHRISTINE L. BORGMAN ET AL., FOSTERING LEARNING IN THE NETWORKED WORLD: THE CYBERLEARNING OPPORTUNITY AND CHALLENGE. A 21ST CENTURY AGENDA FOR THE NATIONAL SCIENCE FOUNDATION. REPORT OF THE NSF TASK FORCE ON CYBERLEARNING (2008), http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf08204 (last visited Jan 1, 2012).

[31] The world's most valuable resource is no longer oil, but data, *supra* note 1.

[32] Sarah Brown, *Where Every Student Is a Potential Data Point*, THE CHRONICLE OF HIGHER EDUCATION, 2017, http://www.chronicle.com/article/Where-Every-Student-Is-a/239712 (last visited May 17, 2017).

profiles incorporate data from social networks to assess algorithmically a student's social connectedness.[33]

Integrating data from multiple sources and systems is a non-trivial matter for reasons of technology, measurement, and inference.[34] The higher education community, via an EDUCAUSE initiative funded by the Gates Foundation, has proposed a "Next Generation Learning Environment" that will provide greater interoperability, and a freer flow of data, between applications that gather data about students.[35]

University Responsibilities for Data

The massive data collection by universities creates vast opportunities for research, teaching, learning, service, outreach, and strategic management. These data collections expose universities to new risks and create responsibilities that may converge and diverge in unexpected ways. Four categories of responsibilities are outlined here: stewardship and governance, and protecting privacy, academic freedom, and intellectual property. As a means to focus this vast territory, the discussion draws out issues that are common to research data and to grey data. Privacy concerns are central to this paper, hence academic freedom and intellectual property are discussed in privacy contexts.

Stewardship and Governance

By collecting data, institutions take on responsibilities for managing those data in the short and long term. Among the many descriptions of these roles, such as sustainability, curation, access, and preservation, "stewardship" has become the overarching term. While the term is used in nuanced ways in the scientific, library, archival, and policy communities, stewardship encompasses a commitment to managing data in ways that they remain findable, accessible, and useful.[36] For some kinds of data, stewardship requires indefinite preservation, while for others,

---

[33] Evan Selinger, *With big data invading campus, universities risk unfairly profiling their students*, CHRISTIAN SCIENCE MONITOR, 2015, http://www.csmonitor.com/World/Passcode/Passcode-Voices/2015/0113/With-big-data-invading-campus-universities-risk-unfairly-profiling-their-students (last visited May 22, 2017).

[34] Franke Kreuter & Roger D. Peng, *Extracting information from big data: Issues of measurement, inference, and linkage*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 257–275 (Julia Lane et al. eds., 1 edition ed. 2014).

[35] MALCOLM BROWN, JOANNE DEHONEY & NANCY MILLICHAP, THE NEXT GENERATION DIGITAL LEARNING ENVIRONMENT: A REPORT ON RESEARCH (2015), https://library.educause.edu/resources/2015/4/the-next-generation-digital-learning-environment-a-report-on-research (last visited Oct 16, 2017).

[36] Mark D. Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*, 3 SCIENTIFIC DATA 160018 (2016), http://www.nature.com/articles/sdata201618 (last visited Apr 23, 2016); MALCOLM BROWN, JOANNE DEHONEY & NANCY MILLICHAP, THE NEXT GENERATION DIGITAL LEARNING ENVIRONMENT: A REPORT ON RESEARCH (2015),

stewardship requires regular cycles of record disposal.[37] However, given the dynamic nature of these data collections, traditional archival approaches to sustaining access to static resources are unlikely to suffice. In an "age of algorithms" where datasets are in constant flux, and can be disaggregated and reaggregated continuously for multiple analytical purposes, new approaches are sorely needed.[38]

While universities have broad responsibilities for stewarding the data they collect, acquire, and hold, some individual persons, offices, committees, or other entities must take specific actions, make investments, and manage the daily operations of data stewardship. Determining which entities have which responsibilities, based on what criteria and policies, is the process of governance. The UC Privacy and Information Security Initiative (PISI), discussed in framing this paper, was among the first to address this process in U.S. higher education. The PISI principles explicitly acknowledge the "distributed nature of information stewardship at UC, where responsibility for privacy and information security resides at every level".[39] Universities are unlikely to appoint "data czars" responsible for all manner of research and grey data. More feasible is for an office or committee to wrangle generalized policies, agreements, and governance mechanisms.

Research Data

Responsibility for research data in universities generally defaults to the researchers who collected those data. These are the individuals who have a vested interest in exploiting and protecting these data. They also are the people who know most about the content and context. Local knowledge is essential to data management, given the vast array of data types, domain expertise, policies, and practices. Along with the benefits of local control come limitations in

---

https://library.educause.edu/resources/2015/4/the-next-generation-digital-learning-environment-a-report-on-research (last visited Oct 16, 2017); Ge Peng et al., *A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets*, 13 DATA SCIENCE JOURNAL 231–252 (2015), dx.doi.org/10.2481/dsj.14-049 (last visited Oct 26, 2017); National Digital Stewardship Alliance, NATIONAL DIGITAL STEWARDSHIP ALLIANCE - DIGITAL LIBRARY FEDERATION (2016), http://ndsa.diglib.org// (last visited Oct 26, 2017); Myron P. Gutmann & Francine D. Berman, STEWARDSHIP GAP PROJECT (2016), http://www.colorado.edu/ibs/cupc/stewardship_gap/ (last visited Aug 6, 2016); Daniel Kleppner et al., *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*, 178 (2009), http://www.nap.edu/catalog.php?record_id=12615.

[37] UCLA Corporate Financial Services, RECORDS RETENTION & DISPOSITION GUIDELINES (2017), https://www.finance.ucla.edu/tax-records/records-management/records-retention-disposition-guidelines (last visited Nov 3, 2017); Special Section on Selection, Appraisal, and Retention of Digital Scientific Data, 3 DATA SCIENCE JOURNAL 191–232 (2004), http://www.jstage.jst.go.jp/browse/dsj/3/0/_contents; Council on Governmental Relations, ACCESS TO AND RETENTION OF RESEARCH DATA: RIGHTS AND RESPONSIBILITIES (2006), www.cogr.edu/viewDoc.cfm?DocID=151536.

[38] Clifford Lynch, *Stewardship in the "Age of Algorithms,"* 22 FIRST MONDAY (2017), http://firstmonday.org/ojs/index.php/fm/article/view/8097 (last visited Dec 5, 2017).

[39] PRIVACY AND INFORMATION SECURITY, *supra* note 3 at 8.

expertise and in continuity. In domains with external funding, graduate students and post-doctoral fellows conduct most data collection and perform most of the management tasks. Students and post-docs often write software code, scripts, and algorithms to analyze those data. While experts in a research domain, rarely are they also experts in data management or software engineering. They perform essential research tasks, but are short-term employees whose cohort is replaced every few years as students graduate, fellowships end, and grant projects are completed.[40]

As papers are submitted for publication and as grant closure looms, many authors and investigators are responsible for releasing associated data. If so, they need to find (and often to fund) ways of sustaining access to their data for some specified number of years after the granting period. The preferred solution is usually to deposit datasets in a data archive or repository, whether organized by discipline, data type, or institution, as these entities tend to have long-term commitments and staff responsible for curation. Archiving of digital research data has been under way for at least 50 years by entities such as the World Data Systems,[41] IQSS,[42] and ICPSR.[43] Some agencies fund research and also fund data archives to sustain access to findings, such as the National Institutes of Health (U.S.) and Economic and Social Research Council (U.K.). Other funding agencies may require universities to maintain their own data archives as a condition of receiving grants.[44] Many public archives, however, are funded by research grants, which limits their ability to make indefinite commitments.

Grey Data

Responsibility for grey data is highly diffuse in universities. Those who collect data may become the stewards of those data, or may pass them to other stewards inside or outside the institution. Among the many data collectors and stewards of grey data are libraries, registrars, undergraduate and graduate divisions, schools and departments, instructional development, individual faculty and staff, and administrators of housing, food services, student stores, and many more. Here too, students and other limited-term staff may have substantial responsibility for day-to-day data collection and management. Many of these data have transient value, but many may be kept indefinitely, whether for potential later use as stores cumulate or because it is often easier to keep them than to invest the labor necessary to discard records selectively.

---

[40] BORGMAN, *supra* note 6.
[41] International Council of Scientific Unions, WORLD DATA SYSTEM: TRUSTED DATA SERVICES FOR GLOBAL SCIENCE (2017), https://www.icsu-wds.org/ (last visited Jul 3, 2017).
[42] Harvard Institute for Quantitative Social Science, (2017), https://www.iq.harvard.edu/home (last visited Oct 23, 2017); IQSS Dataverse Network, (2017), https://dataverse.harvard.edu/ (last visited Oct 22, 2017).
[43] Regents of the University of Michigan, ICPSR - INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (2016), http://www.icpsr.umich.edu/icpsrweb/ICPSR/ (last visited Apr 2, 2013).
[44] EPSRC policy framework on research data: Expectations, U.K. ENGINEERING AND PHYSICAL SCIENCES RESEARCH COUNCIL (2014), https://www.epsrc.ac.uk/about/standards/researchdata/expectations/ (last visited Oct 6, 2017).

Where compliance rules for data protection and management clearly apply, universities will implement those rules. The larger problem is the growing collections of grey data where few rules are explicitly applicable and data stewards must exercise discretion.

Privacy

Privacy is an essential but elusive concept, as Chemerinsky,[45] Solove,[46]Nissenbaum,[47] and others have eloquently explained. It lacks a single core essence, and is best understood as a pluralistic construct that spans information collection, processing, dissemination, accessibility, autonomy, and certain types of intrusion. Privacy is best understood in a context, such as a university's relationship to the data it collects, acquires, and holds. Somewhat different considerations apply to research and to grey data, although even this boundary is porous and mutable.

Privacy issues associated with data usually involve records collected about individuals, which is a foundational area of privacy law and policy. The Code of Fair Information Practice, known as FIPS (or FIPPS for Fair Information Practice Principles), generally applies, regardless of the intended purpose for data collection. FIPS was formulated in the early days of digital records and incorporated in the foundational U.S. laws about government data collection in the 1970s. The U.S. FIPS became the basis for the OECD principles in 1980, updated in 2013, which are widely promulgated and adopted.[48] HIPAA and FERPA, for example, incorporate most of the FIPS principles.

Requirements for notice of data collection and consent to acquire specific kinds of data are the most widely implemented of the FIPS principles. These two principles continue to be required not only in research contexts, but in credit, housing, social media, and any online service that collects data about individuals – even if the notice and consent contract is buried in the fine print of "click through" agreements.[49] Other OECD FIPS principles provide important privacy guidance, such as the Data Quality Principle, which says "personal data should be relevant to the purposes for which they are to be used… and should be accurate, complete, and kept up-to-date;" the Purpose Specification Principle, which requires that the intended uses for collection be specified in advance; and the Use Limitation Principle, that subsequent uses should be limited to

---

[45] Erwin Chemerinsky, *Rediscovering Brandeis's right to privacy*, 45 BRANDEIS LJ 643 (2006), http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/branlaj45&section=28 (last visited Oct 12, 2017).
[46] DANIEL J. SOLOVE, UNDERSTANDING PRIVACY (2/28/10 edition ed. 2010).
[47] HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (1 edition ed. 2009).
[48] THE OECD PRIVACY FRAMEWORK, 1–154 (2013), http://www.oecd.org/internet/ieconomy/privacy-guidelines.htm (last visited Oct 19, 2017); OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA, (1980); RECORDS, COMPUTERS AND THE RIGHTS OF CITIZENS, (1973), https://aspe.hhs.gov/report/records-computers-and-rights-citizens.
[49] CULLEN HOBACK, TERMS AND CONDITIONS MAY APPLY (2013), http://www.imdb.com/title/tt2084953/ (last visited Oct 19, 2017).

those specified and not repurposed without consent of the data subject, unless by other legal authority. Other FIPS principles include security safeguards, openness, individual participation, and accountability.[50]

Protecting privacy by maintaining confidentiality is among the central concerns in human subjects research. Rules for the treatment of human subjects were developed in the same era as the FIPS principles. The Belmont Report established three premises for protection of human subjects: respect for persons, beneficence, and justice. The Belmont principles, in turn, are the basis for Institutional Review Boards (IRB) at universities and other research institutions, which are administered with U.S. government oversight.[51]

Investigators who conduct human subjects research intentionally, as in much of the social sciences, health, and medical domains, will submit their research proposals and protocols to the appropriate Institutional Review Board. The IRB will determine whether the study complies with federal regulations and the amount of oversight required. Some studies are exempt, while others require extensive and continuing review.[52] If human subjects data are to be released upon publication or conclusion of the study, de-identification and anonymization of individuals normally is required, following protocols for best practice in a given domain.

Despite the long history of privacy regulations and best practices in universities, many privacy issues are emerging in areas not clearly covered by FIPS, Institutional Review Boards, or regulations such as HIPAA (medical patient records), FERPA (educational records), and PII (personally identifiable information). These include research projects that capture records of human activity, whether traces of online or offline activity, historical records, or incidental observations of individuals with technologies such as cameras, audio recorders, drones, or other sensors during investigations for other purposes.

Learning analytics are a primary example of grey data that contains sensitive, and often personally identifiable, data about individuals but that is not subject to IRB rules for confidentiality and data protection. Some universities insist on explicit notice and consent to collect data about students' online behavior, but many assume that students have given implicit consent by enrolling in the university. Students may not know what is being collected about them, much less what is being done with those data or who has access to them.[53] FERPA

---

[50] THE OECD PRIVACY FRAMEWORK, *supra* note 48.

[51] THE NATIONAL COMMISSION FOR THE PROTECTION OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH ET AL., BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH. (1979), http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

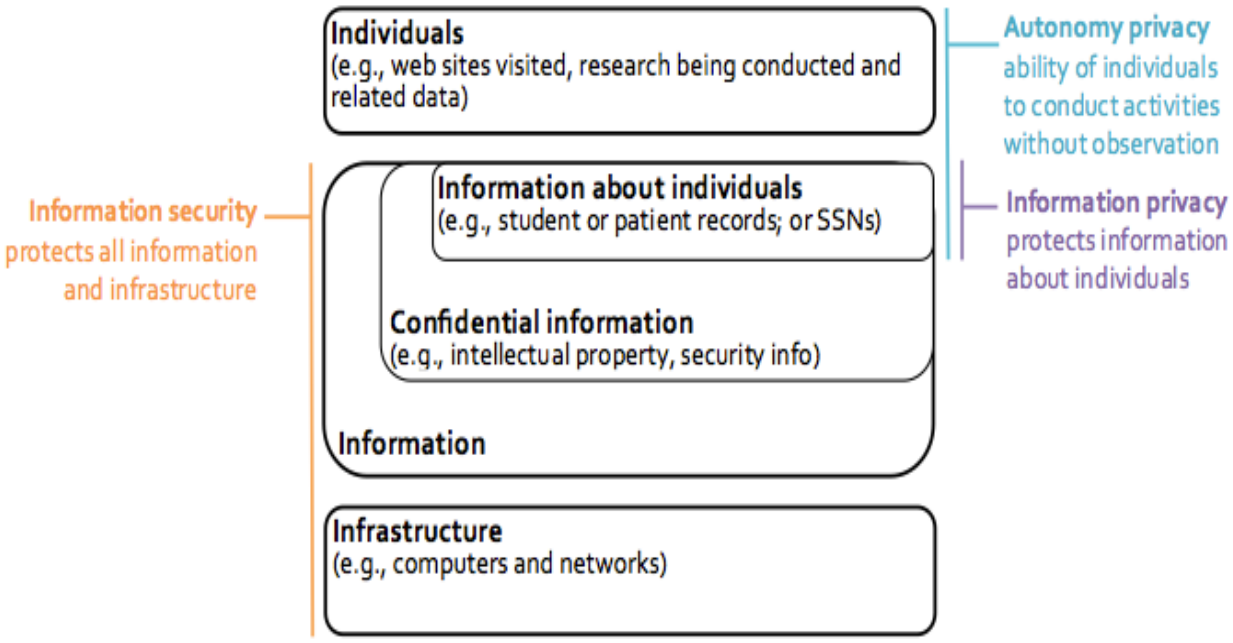
[52] PROPOSED REVISIONS TO THE COMMON RULE: PERSPECTIVES OF SOCIAL AND BEHAVIORAL SCIENTISTS, WORKSHOP SUMMARY, (Robert Pool ed., 2013).

[53] Brown, *supra* note 32; Lisa Ho, NAKED IN THE GARDEN: PRIVACY AND THE NEXT GENERATION DIGITAL LEARNING ENVIRONMENT (2017), https://er.educause.edu:443/articles/2017/7/naked-in-the-garden-privacy-and-the-next-generation-digital-learning-environment (last visited Sep 27, 2017); Asilomar II: Student Data and Records in the Digital Era, (2016),

provides little guidance in using or protecting these data, as learning analytics appear to fall in the generally allowable category of educational uses.[54]

The UC Privacy and Information Security Initiative and the UCLA Data Governance Task Force both addressed data privacy issues by distinguishing between two types of privacy and the security necessary to protect them, as illustrated in Figure 1.

Figure 1: Relationships between autonomy privacy, information privacy, and information security.[55]



Information privacy is narrowly drawn to include specific information about individuals such as those elements in the legal definitions of Personally Identifiable Information (PII). In the

---

https://sites.stanford.edu/asilomar/ (last visited Aug 12, 2017); The Asilomar Convention for Learning Research in Higher Education, (2014), http://asilomar-highered.info/ (last visited Aug 12, 2017); Selinger, *supra* note 33; Sharon Slade & Paul Prinsloo, *Learning analytics: ethical issues and dilemmas*, 57 AMERICAN BEHAVIORAL SCIENTIST 1509–1528 (2013), http://oro.open.ac.uk/36594/ (last visited Oct 6, 2017).

[54] Steven J. McDonald, *A Few Things about E-FERPA*, EDUCAUSEREVIEW, 2013, https://er.educause.edu:443/blogs/2013/1/a-few-things-about-eferpa (last visited Oct 19, 2017); Diana Orrick, *An Examination of Online Privacy Issues for Students of American Universities.*, *in* INTERNATIONAL CONFERENCE ON INTERNET COMPUTING 330–338 (2003), https://www.educause.edu/ir/library/pdf/CSD4039.pdf.

[55] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5; PRIVACY AND INFORMATION SECURITY, *supra* note 3.

California law, PII includes a specific list of data elements,[56] whereas the U.S. federal code is more general: "PII means information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual" (2 CFR 200.79).

Autonomy privacy, or the ability of individuals to conduct activities without observation, is a larger category that subsumes PII. It includes safeguards from surveillance and other kinds of monitoring of behavior. Autonomy privacy overlaps with academic freedom concerns, as discussed in the next section, because it includes the ability to conduct research without being observed. Information security, the third category in the PISI and DGTF reports, protects the confidentiality, integrity, and availability of information, and thus includes the protection of intellectual property. Separating these three concepts led the committees to a broader framing of privacy, security, and governance and to more concise recommendations. These categories are loosely based on legal distinctions between informational and autonomy privacy; security is necessary because current technologies have led to an "unprecedented ability to learn the most intimate and personal things about individuals … (and) unprecedented access to information about individuals".[57]

Academic Freedom

Like privacy, academic freedom is a complex and elusive concept. In considering university responsibilities for data, it intersects with privacy and with freedom of speech. The most succinct, and most widely adopted, statement of academic freedom is that "Teachers are entitled to full freedom in research and in the publication of the results" [58] because academic freedom is "fundamental to the advancement of truth".[59] It is not an absolute right to free speech; rather, the formal statement of academic freedom distinguishes between speech on one's area of expertise and speech as a private citizen, and includes conditions such as adequate performance of other academic duties.[60]

Protecting autonomy privacy is essential to protecting academic freedom. In research contexts, faculty need to be able to protect research in progress, including research data, in the free pursuit of inquiry. Scholars often "test ideas in extreme form" as a means to develop hypotheses,

---

[56] COMMITTEE ON PRIVACY AND CONSUMER PROTECTION, ASSEMBLY BILL NO. 1541 CHAPTER 96. AN ACT TO AMEND SECTION 1798.81.5 OF THE CIVIL CODE, RELATING TO PRIVACY. CIVIL CODE (2015), http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160AB1541.

[57] Chemerinsky, *supra* note 45 at 656.

[58] American Association of University Professors, 1940 STATEMENT OF PRINCIPLES ON ACADEMIC FREEDOM AND TENURE 14 (1940), https://www.aaup.org/report/1940-statement-principles-academic-freedom-and-tenure (last visited Oct 23, 2017).

[59] American Association of University Professors, RESOURCES ON ACADEMIC FREEDOM (2017), https://www.aaup.org/our-programs/academic-freedom/resources-academic-freedom (last visited Oct 5, 2017).

[60] American Association of University Professors, *supra* note 58.

brainstorm with collaborators, or provoke internal debate.[61] Releasing private communications risks mischaracterizing the research and the individuals involved, and thus limits the free pursuit of truth and inquiry. Research data are part of the research process, and thus similarly subject to protection on the grounds of academic freedom and autonomy privacy.[62]

Academic freedom protection is normally associated with academic tenure.[63] Autonomy privacy, however, applies more broadly to the university community. Non-tenured faculty, research staff, and students also conduct research and those data deserve similar protections. Autonomy privacy goes beyond the scope of academic freedom, which covers research and teaching, to include "the ability of individuals to conduct activities without observation".[64] These recent UC initiatives reinforce long-standing university policy on protecting electronic communications and media: "The University recognizes that principles of academic freedom and shared governance, freedom of speech, and privacy hold important implications for the use of electronic communications." … with very limited exceptions, "the University does not examine or disclose electronic communications records without the holder's consent.[65] These are strong protections against electronic surveillance. They also reinforce FIPS, requiring notice and consent to collect data about individuals.

Grey data, such as digital records about teaching and student learning, are similarly covered under the UC Electronic Communications Policy and the adopted recommendations about privacy, information security, and data governance. However, the University of California data and electronic communication policies appear to provide much stronger protections of privacy and academic freedom than are typical of U.S. institutions of higher education.

Intellectual Property

Ownership of intellectual property carries a large set of rights and responsibilities, some which are associated with privacy protection and intrusion. Corporate owners of scholarly publishing, mass media, and social media content deploy "digital rights management" (DRM) technologies to track uses and users in minute detail. These technologies have eroded traditional protections of

---

[61] Joint Senate-Administration Task Force on Academic Freedom, STATEMENT ON THE PRINCIPLES OF SCHOLARLY RESEARCH AND PUBLIC RECORDS REQUESTS UCLA ACADEMIC PERSONNEL OFFICE (2012), https://apo.ucla.edu/policies-forms/academic-freedom (last visited Sep 27, 2017).

[62] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5; PRIVACY AND INFORMATION SECURITY, *supra* note 3.

[63] American Association of University Professors, *supra* note 59; Erwin Chemerinsky, *Is tenure necessary to protect academic freedom?*, 41 AMERICAN BEHAVIORAL SCIENTIST 638–651 (1998), http://journals.sagepub.com/doi/abs/10.1177/0002764298041005005 (last visited Oct 12, 2017).

[64] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5; PRIVACY AND INFORMATION SECURITY, *supra* note 3.

[65] University of California, Office of the President, ELECTRONIC COMMUNICATIONS POLICY 10 (2005), policy.ucop.edu/doc/7000470/ElectronicCommunications.

privacy and intellectual freedom in libraries and other domains.[66] Universities, hospitals, and private businesses who own or control medical patient records are responsible for protecting the confidentiality of those records and limiting their dissemination. Despite regulations, the health industry has found ways to monetize these records, invading privacy and causing other harms to patients.[67] Universities have special responsibilities for managing their intellectual property in ways that protect the privacy of their communities and minimize harm.

Funding agencies usually hold principal investigators responsible for data management plans and other rules associated with intellectual products of research. Journals hold authors responsible for releasing or depositing data, when such rules apply. Scholars acquire many kinds of data over the course of their careers, often at great personal expense. As a consequence of these practices, faculty tend to hold research records, observations, physical samples, and other types of research data as their own property for most intents and purposes. Laboratory notebooks have special status in fields where patent protection may arise.[68] Strictly speaking, many research data may be considered facts and thus not subject to copyright or to ownership.[69] However, the nature of "facts" is a subject of dispute among historians, philosophers, social scientists, and lawyers alike.[70]

---

[66] Julie E. Cohen, *A Right to Read Anonymously: A Closer Look at "Copyright Management" In Cyberspace*, 28 CONNECTICUT LAW REVIEW 981–1039 (1996); Clifford Lynch, *The rise of reading analytics and the emerging calculus of reader privacy in the digital world*, 22 FIRST MONDAY (2017), http://journals.uic.edu/ojs/index.php/fm/article/view/7414 (last visited May 7, 2017).

[67] BEN GOLDACRE, BAD PHARMA: HOW DRUG COMPANIES MISLEAD DOCTORS AND HARM PATIENTS (2012); Patrick Radden Keefe, *The Family That Built an Empire of Pain*, THE NEW YORKER, 2017, https://www.newyorker.com/magazine/2017/10/30/the-family-that-built-an-empire-of-pain (last visited Oct 27, 2017); James F. Peltz & Melody Petersen, *L.A. billionaire cancer doctor Patrick Soon-Shiong battles business turbulence*, LOS ANGELES TIMES, July 5, 2017, http://www.latimes.com/business/la-fi-soon-shiong-20170705-story.html (last visited Oct 27, 2017).

[68] Colin L. Bird, Cerys Willoughby & Jeremy G. Frey, *Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences*, 42 CHEMICAL SOCIETY REVIEWS 8157–8175 (2013), http://pubs.rsc.org/en/Content/ArticleLanding/2013/CS/C3CS60122F (last visited Oct 28, 2017); Jason T. Nickla & Matthew B. Boehm, *Proper Laboratory Notebook Practices: Protecting Your Intellectual Property*, 6 J NEUROIMMUNE PHARMACOL 4–9 (2011), https://link.springer.com/article/10.1007/s11481-010-9237-4 (last visited Oct 28, 2017); Kalpana Shankar, *Order from chaos: The poetics and pragmatics of scientific recordkeeping*, 58 J. AM. SOC. INF. SCI. 1457–1466 (2007), http://onlinelibrary.wiley.com/doi/10.1002/asi.20625/abstract (last visited Oct 28, 2017).

[69] PETER BALDWIN, THE COPYRIGHT WARS: THREE CENTURIES OF TRANS-ATLANTIC BATTLE (2014), http://press.princeton.edu/titles/10303.html.

[70] ANN M. BLAIR, TOO MUCH TO KNOW: MANAGING SCHOLARLY INFORMATION BEFORE THE MODERN AGE (2010); Daniel Rosenberg, *Data before the Fact*, *in* "RAW DATA" IS AN

Although many universities, including the University of California, claim ownership of research data, researchers may be largely unaware of these regulations unless disputes arise or an individual faculty member wishes to take a substantial trove of data to another university when changing jobs.[71] Little guidance exists for how data ownership policies apply to data release requirements. The UC policy cited for data ownership is the last sentence of this paragraph in the Academic Personnel Manual(emphasis added):[72]

> 5. Publicity of Results
> All such research shall be conducted so as to be as generally useful as possible. To this end, the right of publication is reserved by the University. The University may itself publish the material or may authorize, in any specific case, a member or members of the faculty to publish it through some recognized scientific or professional medium of publication. A report detailing the essential data and presenting the final results must be filed with the University. **Notebooks and other original records of the research are the property of the University.**

Given the advances in research practice and digital records since the policy was established in 1958, these issues are receiving renewed attention by the Academic Senate and other UC bodies.

Ownership and responsibility for grey data is particularly problematic. While university records presumably are property of the university, many individuals and units may be involved in data collection, analysis, reporting, and management. As records are mined and combined, tracking sources and policies associated with individual datasets becomes more difficult. In principle, students own the intellectual property in their coursework, such as papers and assignments, yet some of that work and associated online activities may be captured by learning management

---

OXYMORON 15–40 (Lisa Gitelman ed., 2013); "RAW DATA" IS AN OXYMORON, (Lisa Gitelman ed., 2013).

[71] Bradley J. Fikes, *UC San Diego sues USC and scientist, alleging conspiracy to take funding, data - LA Times*, LOS ANGELES TIMES, July 5, 2015, http://www.latimes.com/local/education/la-me-ucsd-lawsuit-20150706-story.html (last visited Jul 8, 2015); Larry Gordon, Gary Robbins & Bradley J. Fikes, *What's behind UCSD, USC court battle?*, SANDIEGOUNIONTRIBUNE.COM, July 9, 2015, http://www.sandiegouniontribune.com/news/science/sdut-usc-ucsd-alzheimers-paul-aisen-court-legal-2015jul19-story.html (last visited Oct 6, 2017); Gary Robbins, *UC San Diego wins legal battle in dispute with USC over Alzheimer's project*, LOS ANGELES TIMES, July 24, 2015, http://www.latimes.com/local/california/la-me-0725-uc-sandiego-20150725-story.html (last visited Oct 6, 2017); Gary Robbins & Bradley J. Fikes, *USC siphons away most of Alzheimer's program*, SANDIEGOUNIONTRIBUNE.COM, August 29, 2015, http://www.sandiegouniontribune.com/news/science/sdut-ucsd-usc-alzheimers-aisen-cooperative-study-2015aug29-htmlstory.html (last visited Oct 6, 2017).

[72] ROBERT G. SPROUL, UNIVERSITY OF CALIFORNIA REGULATION NO. 4 (GENERAL UNIVERSITY POLICY REGARDING ACADEMIC APPOINTEES: SPECIAL SERVICES TO INDIVIDUALS AND ORGANIZATIONS) 7 3 (1958), http://www.ucop.edu/academic-personnel-programs/_files/apm/apm-020.pdf.

systems or other educational technologies. When commercial partners are involved in data collection, either via university contracts or software tools deployed by individual faculty, licensing and ownership of grey data may be unclear or opaque.[73]

The Privacy Frontier

The drive to collect data at ever greater volumes, velocity, and variety is advancing universities toward the privacy frontier at a far faster rate than most administrators, faculty, researchers, or students are aware. Universities are competitive institutions, both internally and externally. Those who exploit data most effectively will gain research grants, awards, students, administrative efficiencies, and other rewards. Those who govern and steward their data most effectively are likely to gain the greatest long-term advantages. On shorter horizons, it is all too easy to exploit data in ways that risk violations of privacy. Protecting privacy adds a layer of complexity to exploiting data, but an essential layer. Institutions ignore privacy at their peril, and the perils are perhaps greatest for universities as guardians of public trust. Technologies tend to advance at a much faster pace than does the law or social practice.[74] When the technologies are in the realm of ideas and knowledge production, as is the case with research and grey data, the stakes for universities are especially high.

Access to Data

Determining who has access to what data, by what criteria, when, and under what conditions is an overarching problem of data governance and stewardship. Competing values often are at stake. Openness promotes transparency and accountability, but can undermine privacy, confidentiality, and anonymity. Confidentiality is essential to protecting human subjects, but can limit the uses of data and the ability to reuse data for other purposes. Trust derives from openness in some situations and confidentiality in others. Long-term stewardship is necessary for longitudinal research and for many kinds of data aggregation, but may result in retaining sensitive records that should be purged regularly by law, policy, or ethical judgment. Access policies that apply to any given data collection may be multiple, conflicting, and change over time.

Privacy concerns abound at the intersection of research and grey data due to the vagaries of defining "research" and "research data." As discussed above, the boundaries of what is considered research are fluid. Materials collected for administrative or teaching purposes may

---

[73] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5.

[74] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999); LAWRENCE LESSIG, THE FUTURE OF IDEAS: THE FATE OF THE COMMONS IN A CONNECTED WORLD (2001); DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS (2017); SOLOVE, *supra* note 46; JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE (2012); ANITA ALLEN & MARC ROTENBERG, PRIVACY LAW AND SOCIETY (3 edition ed. 2015).

later be considered useful for research. Conversely, data collected for research purposes might be put to practical use in university operations later.

One of the major difficulties in implementing policies for open access to research data is the lack of agreement on what content, formats, media, or artifacts are subject to release. Funding agencies and journals generally leave these specifics to investigators, as data may be released in varying states of processing. Rules and practices vary widely by agency and research domain. "Raw" data may be released, with or without sufficient documentation to make them useful to others. Conversely, highly processed data might be released, with or without sufficient documentation, software, and code to make them useful to others. Investigators may meet "the letter of the law" by releasing enough information to satisfy agency or journal requirements, while retaining control over proprietary materials that assure a competitive edge in research. Privacy protection may or may not be an issue, depending on the content of the data.[75]

When disputes arise between researchers, collaborators, funding agencies, or journals about what data are subject to release, universities may need to arbitrate in this unsettled territory. Particularly sensitive, for example, are data from grant projects that constitute dissertation research. To ensure that students can complete their research, that research subjects' confidentiality is protected, and that grant contracts are completed, balancing tests may be necessary. Among the reasons that research data are not released is that specific responsibility for depositing or posting data may be unclear. In most domains, data release is not part of regular scholarly practice. Rarely are the principles or mechanics of data management and dissemination covered in graduate courses on research methods. Graduate students and post-doctoral fellows are the primary data-handlers in most research teams. They may not take, or be given, the data deposit responsibility by principal investigators, for example.[76]

Despite elaborate rules about what constitutes human subjects research, IRBs vary in their judgment of how sensitive any given study may be. For example, IRBs may disagree about necessary protections for records of online activity or historical records. A recent study conducted by researchers at Cornell University and Facebook that manipulated Facebook feeds raised a firestorm of ethical issues in mainstream and social media. A central question raised was when, and to what degree, did the university's IRB review the proposal. The study appears to be

---

[75] Christine L. Borgman et al., *Knowledge infrastructures in science: data, diversity, and digital libraries*, 16 INT J DIGIT LIBR 207–227 (2015), http://link.springer.com/article/10.1007/s00799-015-0157-z (last visited Aug 13, 2015); Christine L. Borgman, Jillian C. Wallis & Matthew S. Mayernik, *Who's Got the Data? Interdependencies in Science and Technology Collaborations*, 21 COMPUT SUPPORTED COOP WORK 485–523 (2012); "RAW DATA" IS AN OXYMORON, *supra* note 70; Pasquetto, Randles, and Borgman, *supra* note 26; Jillian C. Wallis, *The Distribution of Data Management Responsibility within Scientific Research Groups*, 2012, http://escholarship.org/uc/item/46d896fm (last visited Jul 23, 2014).

[76] Wallis, *supra* note 75; Jillian C. Wallis, Elizabeth Rolando & Christine L. Borgman, *If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology*, 8 PLOS ONE e67332 (2013).

legal, per Facebook user agreements; experts disagree about the ethics of using information about individuals in this way.[77]

If an IRB decides that a project does not require IRB review, investigators and staff may have no alternative venue to consult. If sensitive data collection originates outside of the research realm, such as learning analytics, no consultation source may exist beyond the boundaries of system or project. Only if and when someone wishes to publish findings from such studies does an IRB review them, by which time sensitive data may have been collected inappropriately. These data could fall under open access release policies, depending on funding sources and publication venues. UCLA is unusual in providing an alternative consulting entity, which is the Privacy and Data Protection Board. That board is advisory, and consists of faculty and administrators with a broad array of expertise in privacy matters.[78]

A growing concern is sensitive data about individuals collected in the process of technological research that is not submitted for IRB review. Researchers in engineering, for example, may have little experience with human subjects research and be unfamiliar with DHHS and IRB rules. When robotics students test image-recognition algorithms by scattering cameras around a campus, they are likely to capture all manner of human activity without notice or consent of the individuals whose images and actions are recorded. Drones are the current technology of concern, due to their surveillance capabilities and potential for harm to persons and property. Universities are beginning to grapple with ways to balance data protection with innovation in these areas.[79] Technical data such as these could be subject to open access policies, and could inadvertently release data that are subject to PII or other protections.

---

[77] Reed Albergotti & Elizabeth Dwoskin, *Facebook Study Sparks Soul-Searching and Ethical Questions*, WALL STREET JOURNAL, June 30, 2014, http://www.wsj.com/articles/facebook-study-sparks-ethical-questions-1404172292 (last visited Oct 5, 2017); Chris Chambers, *Facebook Fiasco: Was Cornell University's study of 'emotional contagion' a breach of ethics?*, THE GUARDIAN, July 1, 2014, http://www.theguardian.com/science/head-quarters/2014/jul/01/facebook-cornell-study-emotional-contagion-ethics-breach (last visited Oct 5, 2017); Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental evidence of massive-scale emotional contagion through social networks*, 111 PNAS 8788–8790 (2014), http://www.pnas.org/content/111/24/8788 (last visited Oct 5, 2017); Robinson Meyer, *Everything We Know About Facebook's Secret Mood Manipulation Experiment*, THE ATLANTIC, 2014, https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/ (last visited Oct 5, 2017); Gail Sullivan, *Cornell ethics board did not pre-approve Facebook mood manipulation study*, WASHINGTON POST, July 1, 2014, https://www.washingtonpost.com/news/morning-mix/wp/2014/07/01/facebooks-emotional-manipulation-study-was-even-worse-than-you-thought/ (last visited Oct 5, 2017).

[78] UCLA Board on Privacy and Data Protection, UNIVERSITY OF CALIFORNIA LOS ANGELES (2017), http://privacyboard.ucla.edu/ (last visited Oct 29, 2017).

[79] Brandon Stark, UC UNMANNED AIRCRAFT SYSTEM SAFETY | UCOP (2016), http://www.ucop.edu/enterprise-risk-management/resources/centers-of-excellence/unmanned-aircraft-systems-safety.html (last visited Oct 12, 2017).

Uses and Misuses of Data

Among the greatest promises of "big data" is the ability to exploit data for innovative purposes, especially uses that were not anticipated at the time of data collection.[80] Data exploitation can lead to scientific breakthroughs, philosophical insights, and to new products and services. When data exist, clever people will find new uses for those data. Therein lies the rub: how to encourage innovation while protecting against inappropriate, privacy-invading uses of those data? Data systems subject to strict compliance regulations such as IRB, HIPAA, FERPA, and PII may be a declining portion of university data acquisition. The privacy frontier is the vast territory outside those regulated systems.

Anticipating Potential Uses and Misuses

When the Code of Fair Information Practices was developed, nearly 40 years ago, data collection was vastly smaller in scale and information systems were more discrete entities. At today's scale of data collection and aggregation, the original FIPS principles provide much less privacy protection. Revisions of FIPS issued in 2013 by the OECD addressed practical implementations based on risk management and improvements in interoperability of data systems.[81] Notice and informed consent, the foundational FIPS principles, remain necessary but are no longer sufficient.[82] When individuals consent to the collection of specific data elements, they may be giving much broader permissions than anticipated, especially when the stated purposes provide wide latitude for use in research, personalization, improving system performance, or other vagaries.

Broader data collection, for more generic purposes, increases the potential for misuses of data and for privacy risks. The benefits and risks of big data in universities can be balanced by two means. One is to adhere more broadly to the FIPS principles, including collection limitation, data quality, use specification, and purpose specification principles. Common to both FIPS and the tenets of privacy by design is to limit data collection and to state an express justification for each data element to be acquired.[83] The second means is to govern uses of data, once collected.

---

[80] BORGMAN, *supra* note 6; Tom Kalil, BIG DATA IS A BIG DEAL | THE WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY (2012), http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (last visited Apr 1, 2013); ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES (1 edition ed. 2014); PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT, (Julia Lane et al. eds., 1 edition ed. 2014); MAYER-SCHONBERGER AND CUKIER, *supra* note 8.

[81] THE OECD PRIVACY FRAMEWORK, *supra* note 48.

[82] Susan Landau, *Control use of data to protect privacy*, 347 SCIENCE 504–506 (2015), http://www.sciencemag.org/content/347/6221/504 (last visited Aug 7, 2015); UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5; SOLOVE, *supra* note 46.

[83] P.E. Agre, *Institutional circuitry: Thinking about the forms and uses of information*, 14 INFORMATION TECHNOLOGY AND LIBRARIES 225–230 (1995); Bellotti and Sellen, *supra* note 9;

Governance should include specifying who has access to what data, when, and under what circumstances, and identifying what uses are considered appropriate and inappropriate. As criteria for these judgments can change over time, governance processes to assure continuing oversight also are needed.[84]

Individual data elements that appear innocuous at the time of collection can become sensitive in later contexts. In a recent example, students' permanent addresses, which universities maintain in case of emergency, may reveal legal residency status to immigration authorities. Recent changes in the status of Dream Act (DACA) students made this information extremely sensitive.[85] In a much earlier example, universities learned not to expose local home addresses after students were stalked, some fatally.

Potential misuse of research data is a concern often mentioned by those reluctant to release data associated with grants or publications.[86] Data can be taken out of context to make misleading or incorrect inferences, as when health and climate data are used selectively to make claims that run counter to those of the investigators.[87]

Reusing Data

One person's good use or reuse of data may be seen by others as a misuse. The ability to reuse data effectively depends on factors such as the quality of the original data collection, the degree of documentation provided to interpret protocols and context, and the availability of associated software, code, and instrumentation.[88] Whether research data or grey data, problems arise in measurement, because collecting good data is hard to do. Considerable sophistication in the design of research or other protocols is necessary, combined with expertise in statistics and

---

Ann Cavoukian, PRIVACY BY DESIGN: THE 7 FOUNDATIONAL PRINCIPLES (2011), https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf; THE OECD PRIVACY FRAMEWORK, *supra* note 48.

[84] UCLA DATA GOVERNANCE TASK FORCE FINAL REPORT AND RECOMMENDATIONS, *supra* note 5; PRIVACY AND INFORMATION SECURITY, *supra* note 3.

[85] Adam Harris, *Colleges Deplore Trump's Threat to DACA. How Far Can They Go to Fight It?*, THE CHRONICLE OF HIGHER EDUCATION, 2017, http://www.chronicle.com/article/Colleges-Deplore-Trump-s/241110 (last visited Oct 19, 2017).

[86] Wallis, Rolando, and Borgman, *supra* note 76; BORGMAN, *supra* note 6.

[87] Paul N. Edwards, *Global Climate Science, Uncertainty and Politics: Data-laden Models, Model-Filtered Data*, 8 SCIENCE AS CULTURE 437–472 (1999); PAUL N. EDWARDS, A VAST MACHINE: COMPUTER MODELS, CLIMATE DATA, AND THE POLITICS OF GLOBAL WARMING (2010); BEN GOLDACRE, BAD SCIENCE (2008); GOLDACRE, *supra* note 67.

[88] Pasquetto, Randles, and Borgman, *supra* note 26; Matthew S. Mayernik, *Research data and metadata curation as institutional issues*, 67 J ASSN INF SCI TEC 973–993 (2016), http://onlinelibrary.wiley.com/doi/10.1002/asi.23425/abstract (last visited Apr 6, 2016); BORGMAN, *supra* note 6.

methods of data cleaning.[89] Surveys, for example, are far more complex to design, execute, analyze, and interpret than is apparent to the novice researcher – or to the staff member assigned to evaluate a service or system. Problems also arise in interpreting and drawing inferences from data, as much must be known about the purposes and context in which the data were collected.

The potential for misuse and abuse multiply when data elements are aggregated, whether from one data resource or many. Variable names, units of measurement, research protocols, and circumstances of data collection introduce errors that are difficult to assess when combining data. Reliability and validity concerns abound. Estimates of the amount of labor required to "clean" data for aggregation are hard to find; one source suggests devoting about 80% of the work to cleaning and integration.[90] Data science is an inexact science, at best.

Despite these cleaning and analysis problems, data scientists have been remarkably effective at re-identifying individuals by aggregating records from multiple sources.[91] Researchers who wish to use sensitive data about individuals, such as medical records or certain types of surveys, often are required to sign agreements that they will not attempt to re-identify the research subjects.[92]

Intellectual property concerns also arise in aggregating data from multiple sources, whether from research, administrative, or external sources. While any individual dataset may carry documentation about ownership and licensing, maintaining intellectual property information in provenance records through multiple generations of use is proving to be a frontier problem in the data sciences. Despite attaching licenses to datasets that protect privacy, that information can be lost downstream.[93]

---

[89] Kreuter and Peng, *supra* note 34; WILLIAM R SHADISH, THOMAS D. COOK & DONALD T. CAMPBELL, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR GENERALIZED CAUSAL INFERENCE (2002).

[90] MAYER-SCHONBERGER AND CUKIER, *supra* note 8.

[91] Boris Lubarsky, *Re-Identification of "Anonymized" Data*, THE GEORGETOWN LAW TECHNOLOGY REVIEW (2017), https://www.georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/ (last visited Nov 3, 2017); Yves-Alexandre de Montjoye et al., *Unique in the shopping mall: On the reidentifiability of credit card metadata*, 347 SCIENCE 536–539 (2015), http://www.sciencemag.org/content/347/6221/536 (last visited Aug 7, 2015); Latanya Sweeney, *k-anonymity: a Model for Protecting Privacy*, 10 INTERNATIONAL JOURNAL ON UNCERTAINTY, FUZZINESS AND KNOWLEDGE-BASED SYSTEMS 557–70 (2002); Latanya Sweeney, *Matching known patients to health records in Washington State data*, (2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2289850 (last visited May 8, 2017).

[92] Jared A. Lyle, George C. Alter & Ann Green, *Partnering to Curate and Archive Social Science Data*, *in* RESEARCH DATA MANAGEMENT: PRACTICAL STRATEGIES FOR INFORMATION PROFESSIONALS (2014); PROPOSED REVISIONS TO THE COMMON RULE, *supra* note 52.

[93] Chaitanya Baru, *Sharing and caring of eScience data*, 7 INT J DIGIT LIBR 113–116 (2007), https://link.springer.com/article/10.1007/s00799-007-0029-2 (last visited Oct 30, 2017); Jane Hunter & Kwok Cheung, *Provenance Explorer-a graphical interface for constructing scientific publication packages from provenance trails*, 7 INT J DIGIT LIBR 99–107 (2007), https://link.springer.com/article/10.1007/s00799-007-0018-5 (last visited Oct 30, 2017);

Responsibilities for Data Collections

Responsibility for data is particularly diffuse in universities, although similar issues arise in all institutions. Research data collections are scattered across labs and stored on laptops or local servers. Multiple generations of students and staff may have access to these data, which can cumulate over long periods of time. Few of these data may involve human subjects and few of these data may be privacy-sensitive, especially when used alone. Similarly, vast collections of grey data are scattered across universities and cumulated over time. Many are purged regularly on a records-retention cycle, but many are not. Access to campus collections may be limited to the few staff who are certified for their use. In other cases, generations of student workers and other transient labor may use grey data daily in their jobs.

As universities outsource more computing systems and services to commercial entities, they relinquish a substantial degree of control over the data collected by their online systems. When universities purchase licenses for access to digital resources such as publications and grey literature, those contracts may allow data providers to track usage by identifiable individuals, in ways that undermine libraries' abilities to protect traditional rights to read anonymously.[94] Similar problems arise when universities partner with vendors for shared usage of data about individuals, such as analytics on learners or patients, whether for graduation rates or treatment outcomes. Universities are becoming more sophisticated about building privacy and security protections into contracts, especially in cases where vendors have offered to sell universities data about their users.

Yet harder problems arise when faculty or staff require students to use third-party online tools that are not licensed by the university. These "free" online tools are attractive because they offer sophisticated activities, content, or evaluation capabilities suitable for a particular course. However, these tools can collect personal data about their users that are shared with outside partners. Students may have little choice but to opt-in to usage agreements if the software is required for course activities. Instructors and students are often unaware of the privacy and security risks involved. Despite warnings by technology professionals not to install such

---

Mayernik, *supra* note 88; Andrew E. Treloar & Mingfang Wu, *Provenance in support of ANDS' four transformations*, 11 INTERNATIONAL JOURNAL OF DIGITAL CURATION 183 (2016), http://ijdc.net/index.php/ijdc/article/view/11.1.183 (last visited Oct 29, 2017); Michael Wright et al., *Connecting digital libraries to eScience: the future of scientific scholarship*, 7 INTERNATIONAL JOURNAL ON DIGITAL LIBRARIES 1–4 (2007), http://dx.doi.org/10.1007/s00799-007-0030-9 (last visited Mar 18, 2010); Paul Groth et al., *Requirements for Provenance on the Web*, 7 INTERNATIONAL JOURNAL OF DIGITAL CURATION 39–56 (2012), http://ijdc.net/index.php/ijdc/article/view/203 (last visited Apr 29, 2014); JAMES CHENY ET AL., REQUIREMENTS FOR PROVENANCE ON THE WEB (2011), http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements.

[94] Lynch, *supra* note 66; Cohen, *supra* note 66; American Library Association, PRIVACY: AN INTERPRETATION OF THE LIBRARY BILL OF RIGHTS ADVOCACY, LEGISLATION & ISSUES (2006), http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy.

software, usage can be difficult to detect, especially by understaffed tech support offices. A shadow network of risky technology lurks on many campuses.

Public Records Requests

Given the continuing advances toward open access to publications and to data over the last several decades, it is counter-intuitive to place public records requests on the privacy frontier. Public access laws are essential to democratic societies, and university researchers often avail themselves of these laws in gaining access to information/[95] However, these laws are being used in political and frivolous ways that threaten academic freedom and privacy.[96]

Two issues relevant to university data collections come to the fore amidst the great morass of law and policy in this area. One is that the more data that universities collect, the larger the pool of resources subject to public records requests – per the principle, "if you can't protect it, don't collect it." Research data on controversial topics such as climate change, guns, tobacco, and abortion are among the most common records requests.[97] Releasing data and communications about research in progress threatens academic freedom and autonomy privacy. State public records laws vary in the degree to which they allow exceptions for research material.

Grey data also can be requested, such as information on the demographics of the student body, marital status of individuals in an academic department, or email correspondence of individual faculty or administrators.[98] As public records requests to universities have become more sophisticated, so too have the responses of university counsel.[99]

---

[95] JON WIENER, GIMME SOME TRUTH: THE JOHN LENNON FBI FILES (1st edition ed. 2000).
[96] Joint Senate-Administration Task Force on Academic Freedom, *supra* note 61.
[97] Larry Bell, *Michael Mann and the ClimateGate Whitewash: Part One*, FORBES, 2011, https://www.forbes.com/sites/larrybell/2011/06/28/michael-mann-and-the-climategate-whitewash-part-one/ (last visited Oct 19, 2017); Suzanne Goldenberg, *Virginia court rejects sceptic's bid for climate science emails*, THE GUARDIAN, March 2, 2012, http://www.theguardian.com/environment/2012/mar/02/virginia-court-sceptic-access-climate-emails (last visited Oct 20, 2017); Florence Olsen, *Historian resigns after report questions his gun research*, CHRONICLE OF HIGHER EDUCATION (2002); Peter Schmidt, *Dispute Over Climate Scientist's Records Pits Academe Against Media Groups*, THE CHRONICLE OF HIGHER EDUCATION, 2014, http://www.chronicle.com/article/Dispute-Over-Climate/143881 (last visited Oct 20, 2017).
[98] Joint Senate-Administration Task Force on Academic Freedom, *supra* note 61; The University of New Mexico, PUBLIC RECORDS REQUEST | THE UNIVERSITY OF NEW MEXICO (2017), https://publicrecords.unm.edu/ (last visited Oct 19, 2017); The University of Southern Mississippi, PUBLIC RECORDS REQUESTS | THE UNIVERSITY OF SOUTHERN MISSISSIPPI (2017), https://www.usm.edu/university-communications/public-records (last visited Oct 19, 2017); The University of Texas at Austin, OPEN RECORDS REQUESTS | FINANCIALS | THE UNIVERSITY OF TEXAS AT AUSTIN (2017), https://financials.utexas.edu/resources/open-records-requests (last visited Oct 19, 2017); University of Washington, REQUEST A PUBLIC RECORD | PUBLIC RECORDS

The second issue is that state public records act requests in the U.S. apply to public universities but not to private universities or corporations. Faculty, students, and staff at public universities thus carry a higher burden in managing their data and in responding to public records requests. Responding to such requests can be extremely time-consuming and expensive, in addition to the risks to academic freedom and privacy. Researchers at public and private universities frequently collaborate with each other, which can expose the data of private universities to these requests. As a result, members of public universities may seek protections of their research materials and communications comparable to those at private universities, which also protects collaborations.[100]

An emerging area of concern is whether trends toward open access to data in some fields may undermine a university's ability to protect data from public records requests in other fields. In some domains of the biosciences, physical sciences, and social sciences, open data is the default condition at the time of publishing research. Some researchers in some domains attempt to work completely in the open, releasing data continuously. In most academic disciplines, however, researchers maintain control of their data and records indefinitely.[101]

Cyber Risk and Data Breaches

Universities are the third highest sector for data breaches, constituting about 10% of reported breaches; healthcare and retail are the top two sectors.[102] From 2005 to late-2017, colleges and universities reported 798 breaches, affecting more than 25 million records.[103][104] Institutions of higher education have extensive data resources and may be perceived as more vulnerable to attack than hospitals, banks, governments, retail, or other entities. Research universities are commonly targeted for the intellectual property manifest in research content. Those with medical centers are targeted for patient records, which are valuable resources for identity and insurance theft. Student records have become high value targets because logon credentials provide access

---

AND OPEN MEETINGS (2017), http://www.washington.edu/publicrecords/request-a-public-record/ (last visited Oct 19, 2017).

[99] John C. Dowling, LETTER FROM UW-MADISON LEGAL COUNSEL REGARDING CRONON EMAILS (2011), http://news.wisc.edu/letter-from-uw-madison-legal-counsel-regarding-cronon-emails/ (last visited Oct 19, 2017); Anthony Grafton, *Wisconsin: The Cronon Affair*, THE NEW YORKER, 2011, https://www.newyorker.com/news/news-desk/wisconsin-the-cronon-affair (last visited Oct 19, 2017); Sara Hebel, *Wisconsin-Madison to Release Professor's E-Mails but Withhold Those Said to Be Private*, THE CHRONICLE OF HIGHER EDUCATION, 2011, http://www.chronicle.com/article/Wisconsin-Madison-to-Release/126994 (last visited Oct 19, 2017).

[100] Joint Senate-Administration Task Force on Academic Freedom, *supra* note 61.

[101] BORGMAN, *supra* note 6.

[102] ISTR20: INTERNET SECURITY THREAT REPORT, (2015).

[103] As of November 3, 2017

[104] Privacy Rights Clearinghouse, DATA BREACHES PRC (2017), https://www.privacyrights.org/data-breaches (last visited Oct 19, 2017).

to expensive licensed content from publishers and other sources. Intruders seeking one kind of information may wander through other databases along the way. Data on individuals that are held by third parties, such as collaborating universities or outside contractors, also are vulnerable to breach.

Universities' challenges in balancing access and protection are among the hardest of any institutional sector. Universities are heterogeneous institutions that acquire many kinds of data and need sophisticated, layered approaches to cyber security. Whereas the financial and intelligence sectors, for example, may prioritize cyber risk protection in the extreme, universities are open by design, encouraging the free flow of information throughout their communities. Individuals partner with collaborators from other institutions, countries, and cultures, which requires shared access to online resources. Campus visitors are vast in number and need access to networks to participate in local activities. Student and staff turnover is high due to short courses and short-term contracts. As a result of these operating conditions, universities must secure their systems and networks without crippling their missions of research, teaching, and service. Research data must flow to students for use in class projects, albeit in a controlled manner. Network security must not become ubiquitous surveillance, lest other harms result.

Cyber risk takes many forms, such as phishing attacks on individuals, viruses, bots, ransomware, data breaches, and distributed denial of service attacks. The list grows by the day. Some risks are obvious, such as the need for many layers of protection on patient data. Others are less obvious, such as attacking a student admissions database for competitive information. Systems are only as well protected as their weakest link. The Target Store breach of credit card records resulted from a successful hack of their HVAC system.[105] A distributed denial of service attack on Netflix was launched by mobilizing networked baby monitors.[106] The ability to mobilize small devices for big attacks will grow as the Internet of Things expands, potentially becoming the "Internet of Terror".[107]

Universities are following the lead of the public and private sectors in enhancing security of their systems, training their communities, and promoting good practices for "cyber health." Deleterious computer-related events are difficult to anticipate and no sector of the economy is immune to attack.[108] No online system ever can be completely secure, any more than any

---

[105] Jaikumar Vijayan, *Target attack shows danger of remotely accessible HVAC systems | Computerworld*, COMPUTERWORLD, 2014, https://www.computerworld.com/article/2487452/cybercrime-hacking/target-attack-shows-danger-of-remotely-accessible-hvac-systems.html (last visited Oct 19, 2017).

[106] Haley Sweetland Edwards, HOW WEB CAMS HELPED BRING DOWN THE INTERNET, BRIEFLY TIME (2016), http://time.com/4542600/internet-outage-web-cams-hackers/.

[107] George V. Neville-Neil, *IoT: The Internet of Terror*, 60 COMMUNICATIONS OF THE ACM 36–37 (2017), https://cacm.acm.org/magazines/2017/10/221328-iot/fulltext (last visited Oct 10, 2017).

[108] Peter G. Neumann, *Far-sighted Thinking About Deleterious Computer-related Events*, 58 COMMUNICATIONS OF THE ASSOCIATION FOR COMPUTING MACHINERY 30–33 (2015), http://doi.acm.org/10.1145/2700366 (last visited Dec 8, 2017); Taylor Armerding, THE 16

building is completely secure from physical attack. By analogy, security comes in layers of locks, cameras, sensors, and alerts.[109] Resilience and recovery also have become watchwords for cybersecurity. The severity of attacks must be minimized, but backup and recovery plans also are necessary.[110] The costs and benefits of each tactic must be evaluated, lest funds spent on protection lessen the investment in the mission of the institution.

A looming challenge on the privacy frontier is how to secure the privacy of human subjects once data are collected. IRBs focus on the design of studies, confidentiality, notice and consent, and good practices for data storage and backup. Their membership is drawn from researchers across campus who have expertise in research design and methods. IRBs, and the university staff that support them, are not necessarily experts in security, cyber risk, cryptography, or in the open data policies to which research projects may be subject. Investigators are required to report on research progress at regular intervals. However, short of known data breaches, IRBs have few mechanisms to follow up on data security. Data management practices vary widely by domain, thus IRBs lack common standards to enforce across campuses.[111] Governance models need to promote more engagement between IRBs, investigators, cyber security units, and other parts of the research enterprise.

BIGGEST DATA BREACHES OF THE 21ST CENTURY CSO ONLINE (2017), https://www.csoonline.com/article/2130877/data-breach/the-16-biggest-data-breaches-of-the-21st-century.html (last visited Sep 22, 2017); Waqas Amir, UNPROTECTED S3 CLOUD BUCKET EXPOSED 100GB OF CLASSIFIED NSA DATA HACKREAD (2017), https://www.hackread.com/unprotected-s3-cloud-bucket-exposed-100gb-of-classified-nsa-data/ (last visited Dec 1, 2017); David Greene, NSA'S HACKERS WERE THEMSELVES HACKED IN MAJOR CYBERSECURITY BREACH NPR.ORG (2017), https://www.npr.org/2017/11/14/564006460/nsas-hackers-are-hacked-in-major-cybersecurity-breach (last visited Nov 14, 2017); Andy Greenberg, HE PERFECTED A PASSWORD-HACKING TOOL—THEN THE RUSSIANS CAME CALLING WIRED (2017), https://www.wired.com/story/how-mimikatz-became-go-to-hacker-tool/ (last visited Nov 13, 2017); Julie Angwin, HOW JOURNALISTS FOUGHT BACK AGAINST CRIPPLING EMAIL BOMBS WIRED (2017), https://www.wired.com/story/how-journalists-fought-back-against-crippling-email-bombs/ (last visited Nov 13, 2017); Susan Landau, *The real security issues of the iPhone case*, 352 SCIENCE 1398–1399 (2016), http://science.sciencemag.org/content/352/6292/1398 (last visited Nov 21, 2017).

[109] BRUCE SCHNEIER, SECRETS & LIES: DIGITAL SECURITY IN A NETWORKED WORLD (2000).

[110] Matthew Goche & William Goueveia, WHY CYBER SECURITY IS NOT ENOUGH: YOU NEED CYBER RESILIENCE FORBES (2014), https://www.forbes.com/sites/sungardas/2014/01/15/why-cyber-security-is-not-enough-you-need-cyber-resilience/ (last visited Oct 12, 2017); IGOR MIKOLIC-TORREIRA ET AL., A FRAMEWORK FOR EXPLORING CYBERSECURITY POLICY OPTIONS (2016), https://www.rand.org/pubs/research_reports/RR1700.html (last visited Sep 27, 2017); National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity*, 1–39 (2014), https://www.cslawreport.com/files/2015/04/07/nist-combined-file.pdf (last visited Oct 12, 2017).

[111] Shankar, *supra* note 68; Melissa H Cragin & Kalpana Shankar, *Scientific data collections and distributed collective practice*, 15 COMPUTER SUPPORTED COOPERATIVE WORK 185–204 (2006).

Curating Data for Privacy Protection

Data management is an expensive endeavor, and one that has come to the fore in the research data arena.[112] Any entity that collects data must make conscious decisions about which data are worth sustaining, which can be discarded, and which might be allowed to fade away.[113] Curating data in ways that maintain privacy protections and reduce risks adds an extra layer of management, and is thus a frontier issue to be addressed.

Digital data do not survive by benign neglect. Computers and websites are replaced; disks crash; software is updated, is no longer available, or is not supported; computer ports and drivers are not compatible with current equipment; data processing pipelines are poorly documented; and those with expertise have graduated or left the university. Digital data remain useful only through investments in curation, documentation, and migration to new formats and systems. Systems and data collections need to be assessed on a cyclical basis, purging sensitive data based on retention rules and refreshing data collections deemed worthy of continuing access. Maintaining provenance records is essential, lest data collections be separated from information about origins; licensing and ownership; applicable regulations; records of notice, consent, and acceptable uses; authorizations for access; and other contexts.[114] Archivists, records managers, and librarians should be closely involved in these processes.

---

[112] FRANCINE BERMAN ET AL., SUSTAINABLE ECONOMICS FOR A DIGITAL PLANET: ENSURING LONG-TERM ACCESS TO DIGITAL INFORMATION (2010), http://brtf.sdsc.edu/publications.html; Berman and Cerf, *supra* note 13; BORGMAN, *supra* note 6.

[113] Christine L Borgman, *Not Fade Away: Social science research in the digital era*, PARAMETERS: SOCIAL SCIENCE RESEARCH COUNCIL, 2016, http://parameters.ssrc.org/2016/06/not-fade-away-social-science-research-in-the-digital-era/ (last visited Jul 14, 2016).

[114] Miriam Ney, Guy K. Kloss & Andreas Schreiber, *Using Provenance to support Good Laboratory Practice in Grid Environments*, ARXIV:1112.3062 [CS] (2011), http://arxiv.org/abs/1112.3062 (last visited Oct 29, 2017); Lucian Carata et al., *A primer on provenance*, 57 COMMUNICATIONS OF THE ACM 52–60 (2014), http://cacm.acm.org/magazines/2014/5/174341-a-primer-on-provenance/abstract (last visited May 19, 2014); Jinfang Niu, *Provenance: crossing boundaries*, 41 ARCHIVES AND MANUSCRIPTS 105–115 (2013), http://www.tandfonline.com/doi/abs/10.1080/01576895.2013.811426 (last visited Mar 31, 2014); IAN FOSTER & LUC MOREAU, PROVENANCE AND ANNOTATION OF DATA (2006), http://www.w3.org/2011/prov/wiki/Connection_Task_Force_Informal_Report; P. Buneman, *Characterizing data provenance*, 1832 *in* ADVANCES IN DATABASES 171–171 (2000), ://000165335600012; Clifford A. Lynch, *When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web*, 52 JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE & TECHNOLOGY 12–17 (2001); Jun Zhao et al., *Linked data and provenance in biological data webs*, 10 BRIEF BIOINFORM 139–152 (2009), http://bib.oxfordjournals.org/content/10/2/139 (last visited Jul 22, 2013).

Responsibility for data collections is highly distributed in universities, which complicates curating data collections in the short and long term. Rare is the researcher with expertise in data management or with the resources to invest in long-term sustainability of research data. Even more rare is the researcher with expertise in data archiving, records management, and the legal vagaries of records retention cycles. Similarly, few of the administrative staff involved in collecting and analyzing grey data are records management experts. All of these individuals and offices need somewhere to turn for guidance and responsibility to ensure that universities make wise choices for what to keep, what to discard, how, and when.

Institutions more readily claim ownership of data than take responsibility for curating those data. Ownership and stewardship need to be more tightly coupled in universities, and probably in most others types of institutions.

Conclusions and Recommendations

Universities are as enamored of "big data" as other sectors of the economy and are similarly effective in exploiting those data to competitive advantage. They have privileged access to research data and to data about their communities, all of which can be mined and combined in innovative ways. Universities also have a privileged social status as guardians of the public trust, which carries additional responsibilities in protecting privacy, academic and intellectual freedom, and intellectual property. Being good stewards of the data with which they are entrusted is especially difficult when conflicts arise between community practices and values. For some kinds of data, good stewardship requires that access to data be sustained indefinitely, and in ways that those data can be reused for new purposes. For other kinds of data, good stewardship requires that they be protected securely for limited periods of time and then destroyed. Criteria for data protection and access can change over time, whether due to different uses of a data collection, such as grey data being mined for research or research data being deployed for operations; transfer of stewardship within and between institutions; changes in laws and policies; or new externalities.

The rate of data collection has grown exponentially over the last decade, both research and grey data within universities, and along with data collection in the other economic sectors with which universities partner. These include government and business, social media, sensor networks, the Internet of Things, and much more. As the ability to mine and combine data improves, and as technologies become more interoperable, the boundaries between data types and origins continue to blur. Responsibilities for stewardship and exposure to cyber risk increases accordingly. Risks to privacy invasion, both information privacy and autonomy privacy, accelerate as most of these data can be associated with individuals, whether as content or creators of data. Anonymity, which is fundamental to most methods of privacy protection, has become extremely difficult to sustain as methods of re-identifying individuals become more sophisticated. Notice and informed consent remain necessary, but are far from sufficient for maintaining privacy in universities or in other sectors.

Open access to publications and to data are social policies that promote transparency and accountability in the research enterprise. Adoption is uneven because costs, benefits, and incentives for open access, especially to data, are aligned in only a few fields and domains. For

most researchers, releasing data involves considerable costs, with benefits going to others. Individual researchers or their employers may bear the costs of data stewardship, whether curation for sustainability or curation for cyclical disposal, depending on which is more appropriate for protecting privacy, academic and intellectual freedom, intellectual property, and other values.

None of these frontier challenges is easily addressed, nor will appropriate responses be consistent across the university sector in the U.S., much less in other countries and cultures. Data are valuable institutional assets, but they come at a price. When individuals and institutions collect data, they must be prepared to protect them. These recommendations, which draw heavily on experiences in the University of California, are offered as starting points for discussion.

Begin with First Principles

Universities should focus on their core missions of teaching, research, and services to address priorities for data collection and stewardship. Tenets of privacy by design, the Code of Fair Information Practice, the Belmont Principles, and codifications of academic and intellectual freedom are established and tested. Implementation is often incomplete, however. For faculty, students, staff, research subjects, patients, and other members of the university community to enjoy protection of information and autonomy privacy, more comprehensive enforcement of principles such as limiting data collection, ensuring data quality, and constraining the purposes for each data element is necessary. Digital data do not survive by benign neglect, nor are records destroyed by benign neglect. Active management is necessary. Notice and consent should never be implicit. When institutions ask for permission to acquire personal data, and are transparent and accountable for uses of data, they are more likely to gain respect in the court of public opinion.

Embed the Ethic

Data practices, privacy, academic and intellectual freedom, intellectual property, trust, and stewardship all are moving targets. Principles live longer than do the practices necessary to implement those principles. Universities are embedding data science and computational thinking into their curricula at all levels. This is an opportune moment to embed data management, privacy, and information security into teaching and practice as well. By encouraging each individual to focus on uses of data, the problem becomes personal. Rather than collecting all data that could conceivably be collected, and exploiting those data in all conceivable ways, encourage people to take a reflective step backwards. Consider the consequences of data collection about oneself, and how those data could be used independently or when aggregated with other data, now and far into the future. Think about potential opportunities and risks, for whom, and for how long. Study data management processes at all levels and develop best practices. Collect the data that matter, not just data that are easy to gather. Interesting conversations should ensue. The Golden Rule still rules.

Promote Joint Governance

The successes of the University of California in developing effective principles for governing privacy and information security have resulted from extensive deliberations between faculty, administrators, and students. These can be long and arduous conversations to reach consensus, but have proven constructive at creating communication channels and building trust. Many years of conversations about information technology policy at UCLA, for example, have resulted in much deeper understanding between parties. Faculty have learned to appreciate the challenges faced by administrators who need to balance competing interests, keep systems running, and pay for infrastructure out of fluctuating annual budgets. Administrators, in turn, have learned to appreciate the challenges faced by faculty who have obligations to collaborators, funding agencies, and other partners scattered around the world, and daily obligations to support students who have disparate skills and access to disparate technologies. Institutional learning is passed down through generations of faculty, students, and administrators through joint governance processes. These mechanisms are far from perfect, and can be slow to respond at the pace of technological change. However, echoing Churchill's assessment of democracy, it works better than any other system attempted to date.

Promote Awareness and Transparency

The massive data breaches of Equifax, Target stores, J.P. Morgan Chase, Yahoo, the National Security Agency, and others have raised community awareness about data tracking, uses of those data by third parties, and the potential for exposure.[115] This is an ideal time to get the community's attention about opportunities and risks inherent in data of all kinds. Individuals, as well as institutions, need to learn how to protect themselves and where to place trust online. When people suspect that personal data are being collected without notice and consent, or when they think they are being surveilled without their knowledge, they may react in anger.[116] Universities are at no less cyber risk than other sectors, but are held to higher standards for the public trust. They have much to lose when that trust is undermined.

Do Not Panic

Panic makes people risk-averse, which is counterproductive. Locking down all data lest they be released under open access regulations, public records requests, or breaches will block innovation and the ability to make good use of research data or grey data. The opportunities in exploiting data are only now becoming understood. Balanced approaches to innovation, privacy,

---

[115] Armerding, *supra* note 108; Amir, *supra* note 108; Greene, *supra* note 108.
[116] Steve Lohr, *At Berkeley, a New Digital Privacy Protest*, THE NEW YORK TIMES, February 1, 2016, https://www.nytimes.com/2016/02/02/technology/at-uc-berkeley-a-new-digital-privacy-protest.html (last visited Oct 30, 2017); The Associated Press, *Online Attacks at UCLA Health Exposed 4.5. Million*, THE NEW YORK TIMES, July 17, 2015, https://www.nytimes.com/2015/07/18/business/online-attacks-at-ucla-health-exposed-4-5-million.html (last visited Oct 30, 2017).

academic and intellectual freedom, and intellectual property are in short supply. Patience and broad consultation of stakeholders is needed.

Acknowledgements

Full disclosure: The author is actively engaged in the University of California activities mentioned herein. She was a founding member of the UCLA Privacy and Data Protection Board, a member of the PISI Steering Committee, Co-Chair of the UCLA Data Governance Task Force, and currently is Chair of the University of California Academic Computing and Communications Committee (UCACC) (2017-2018 academic year; Vice Chair 2015-2017). In her role as a UCACC officer, she is a member of the UC Office of the President Cyber Risk Governance Committee (2015-2018). She has been a member of the Advisory Board to EPIC[117] since its founding in 1994 and served on the EPIC Board of Directors from 2010 to 2017. The opinions in this paper and talk are her own.

Acknowledgements are due to the many colleagues in the University of California who have aided, abetted, and supported these privacy initiatives: Kent Wada, Jim Davis, Amy Blum, Scott Waugh, Dana Cuff, Jerry Kang, Leah Lievrouw, Jan Reiff, David Kay, Maryann Martone, Joanne Miller, Jim Chalfant, Shane White, Sheryl Vacca, Gene Lucas, and other members of the PISI and DGTF committees. My research group at the UCLA Center for Knowledge Infrastructures provided essential critique and commentary on the paper and talk: Irene Pasquetto, Bernie Boscoe, Milena Golshan, Peter Darch, and Michael Scroggins. Morgan Wofford provided extensive bibliographic research. James Dempsey of the Berkeley Center for Law and Technology provided detailed comments on the draft paper. Outside UC, credit is due to Marc Rotenberg and the staff at the Electronic Privacy Information Center; Anne Washington of George Mason University. Special thanks are due to Paul Schwarz, Jim Dempsey, Chris Jay Hoofnagle, and others at the Berkeley Center for Law and Technology, whose invitation to give the Tenth Annual BCLT Privacy Lecture provided the incentive to write this paper, and to Erwin Chemerinsky (Berkeley) and Katie Shilton (University of Maryland) who provided extensive and insightful commentary as respondents to the public lecture on November 16, 2017.

[117] EPIC - Electronic Privacy Information Center, (2017), https://www.epic.org/ (last visited May 8, 2017).