Council Meeting Minutes
March 12-13, 2020

Attendees

Council Members: Dave Armstrong, Bobray Bordelon, Jon Cawthorne, Lisa Cook (Chair), James Doiron, Kristin Eschenfelder, Mark Hansen, Trevon Logan, Lindsey Malcom-Piqueux, Ken Smith, Katherine Wallman, and Esther Wilder

Guests: Barbara Downs (U.S. Census Bureau), David Grimm (University of Michigan Office of the General Counsel), H.V. Jagadish (University of Michigan's Michigan Institute for Data Science), Yajuan Si (University of Michigan Survey Research Center, Timothy Mulcahy (Westat), Simson L Garfinkel (U.S. Census Bureau)

ICPSR Staff: Kehinde Adeniyi, Annahita Akbarifard, Dharma Akmon, Trent Alexander, JD Alford, George Alter, Zachary Bennett, Zachary Bennett, Ashok Bhargav, David Bleckley, Sara Britt, Jon Brode, Sarah Burchart, Scott Campbell, Stephanie Carpenter, Robert Choate, Alina Conn, Gin Corden, Farrah Cundiff , Edward Czilli, Linda Detterman, Ren Dickson, Amanda Draft, Benjmin Dreyer, Alexandra Eastman, Allyson Flaster, Aubrey Garman, Libby Hemphill, Lynette Hoelter, Rachel Huang, Stuart Hutchings, Samuel Imbody, Sanda Ionescu, Abay Israel, Susan Jekielek, Jeffrey Jones, Kevin Kapalla, Lisa Kelley, Kilsang Kim, Jennifer Koski, Piotr Krzystek, Kathryn Lavender, John Lemmer, Susan Leonard, Margaret Levenstein, Scott Liening, Daphne Lin, Jared Lyle, John Marcotte, Trisha Kunst Martinez, Arun Mathur, James McNally, Erin Meyer, A.J. Million, Bianca Monzon, Elizabeth Moss, Justin Noble, Anna Ovchinnikova, Michelle Overholser, Shelly Petrinko, Amy Pienta, Darleen Poisson, Eszter Palvolgyi-Polyak, Daniel Pritts, Tamara Qawasmeh, Raghunath Ravi , Shane Redman, Sarah Rush, Karunakara Seelam, Bing She, Michael Shove, Sarah Snyder, Fillippo Stargell, Chelsea Samples-Steele, Huda Syed , Sandra Tang, David Thomas, Michael Traugott, Allison Tyler, Rujuta Umarji, Vanessa Unkeless-Perez, Diane Viebahn, Xiaosen Wang, Jay Winkler, and LingLing Zhang

Approval of the Minutes

Council Chair Lisa Cook called the meeting to order and asked for approval of the October 2019 Council minutes. The October 2019 minutes were approved unanimously.

Director's Update

Maggie Levenstein, ICPSR Director, gave a presentation on the state of ICPSR. Levenstein reviewed ICPSR's strategic priorities. ICPSR is currently recruiting staff for Lead Architect in CNS, Education data archive assistant research scientist, and a post-doc for MICA (Measuring the Impact of Curation). ICPSR is also hiring in its Curation, CNS, and PUMUS units. Levenstein reviewed the projects that have recently been funded as well as ones that are currently submitted and under review by their sponsors. Levenstein discussed two new relatively new projects: the *Registry of Efficacy and Effective Studies* and *ResearchDataGov*. Levenstein discussed our COA3D proposal to NSF, including what the project is and where it is in the proposal review process. Levenstein introduced Mike Traugott, interim director of the Summer Program. The Summer Program is facing unprecedented challenges with COVID-19. Levenstein introduced Johanna Davidson Bleckman as the new ICPSR Privacy Officer.

Education Committee

Council Attendees: Dave Armstrong, James Doiron, Trevon Logan, Lindsey Malcom-Piqueux (Chair), Katherine Wallman, and Esther Wilder

ICPSR Staff: Scott Campbell, Stephanie Carpenter, Edward Czilli, Linda Detterman, Lynette Hoelter, John Lemmer, Maggie Levenstein, Fillippo Stargell, and Mike Traugott

Mike Traugott (Interim Director of ICPSR Summer Program) began the meeting by discussing his experiences and observations since taking over as Interim Director of the Summer Program.

All program staff are working actively and collaboratively to address the issues that the Summer Program is facing given uncertainty of COVID-19. Several contingency plans have been developed and program staff are actively preparing for alternative arrangements. The Summer Program staff have been meeting daily to plan for all contingencies. They have created a plan and are beginning to implement it. They are prepared to discuss it with Council as well.

Traugott lauded the devoted and knowledgeable program staff. Much of January was spent understanding the type of program data was available, and how these data had been used in the past to understand program outcomes, effectiveness, etc. Going forward, the plan is to collect more robust data where possible and to analyze these data on instructors, personal characteristics, salaries to ensure high quality instruction and equitable treatment across instructors.

Traugott also described steps to be taken to update administrative functions in terms of how the program approaches hands-on work (e.g., dealing with refunds, etc.). Traugott and Levenstein mentioned that they have tried to use commercially available software used for organizing conferences to facilitate many of these tasks.

The work of the Summer Program involves a lot of effort on ICPSR's part. ICPSR also works closely with other institutions to recruit participants and in the provision of financial support for students. For example, the Summer Program is currently working with the American Political Science Association on a program to recruit Middle East and North African students, and the American Political Science Association Political Methodology group to support and recruit women participants to the Summer Program.

The report in the Council Binder details some of the goals for the year, much of which still holds despite the uncertainty caused by the COVID-19 pandemic. This year is the first time that all Teaching Assistant (TA) positions were advertised to create a more diverse group of TAs in the program.

Data regarding diversity among student program participants was discussed and is currently being compiled. Levenstein added that many of the program records have been digitized. The University's Development Office has the data on program alumni, are in process of linking to the other databases at UM. The Summer Program will be getting the report back soon and could use the information to identify potential donors.

Traugott discussed efforts to analyze student evaluation data. Some of this work has been completed by Council Member Dave Armstrong. Armstrong explained that in the past evaluation surveys were done on paper instead of online because there were no records of who was in what class. The schedule that students signed up with on the first day was not necessarily the schedule that they ended up because students move around sections/classes. Acting within the constraint of having pen/paper system, Armstrong tried to automate it as much as possible using OCR software. This reduced the time people were working on the data, and was less error-prone, but not perfect.

Malcom-Piqueux asked whether there has been any discussion on learning outcomes assessment within the Summer Program. Traugott indicated that there is a Center for Research on Learning and Teaching (CRLT) on campus and they have been in discussion with them about how to move enrollments in the Summer Program to Canvas. Summer Program staff can consult with them on questions of assessment. Perhaps the Summer Program can select a subset of the CRLT items for a course or set of courses and take advantage of something that is already in place.

Status of the Summer Program in Light of COVID-19 Pandemic

Traugott updated the committee on the current plans and contingencies for this summer's Program considering COVID-19. In addition to the courses held at the University of Michigan, there are seven off-site locations for workshops, including overseas. If the ICPSR Summer Program cancels the course, students will receive a complete refund. If the Summer Program

offers the course online, registered participants can drop by a certain date and get a refund. If a registered participant wishes to cancel after that date, they can still receive a refund minus a fee of $250.00.

The situation is complex due to the uncertainty and the number of moving parts. The Summer Program must work with instructors and site coordinators for the seven locations. Czilli indicated that the Summer Program will be offering instructors with the tools that the University of Michigan has made available to support teaching with Canvas. The Summer Program staff will recommend a suite of applications for faculty to use for their online instruction and the Summer Program will provide training for participants and faculty not familiar with that application or platform. The goal is to keep people from unilaterally deciding which software or tool they will use for online courses. Currently developing a limited list of delivery mechanisms.

The Committee discussed the Blalock Lectures in an online format. Usually the Summer Program bring scholars to UM for a couple of days where they give 1 to 3 lectures to program participants. The Summer Program wants to expand the Blalock Lecture offerings to include a track for data stewardship. This track might focus on activities like IRB, pre-registration of studies, data management to protect confidentiality, etc.

The effect of these changes on program budget were discussed. Currently, the Summer Program staff cannot predict the full extent of the consequences of the shift to a remote delivery model on revenue, etc. But the experience this summer will certainly inform future considerations of various delivery models. Levenstein added that there is a chance that these changes might draw people who have families that ordinarily limit their mobility to now participate online.

Council Member Esther Wilder raised the issue of the participants in the Diversity Scholarship program in the online model. A latent function of the Summer Program is the network building and the accrual of social capital. If moving to an online environment, it is important to think about what tools could be used to try to re-create network building in online environment.

Levenstein agreed and responded that the Summer Program staff has been thinking about this and is discussing in their planning meeting that morning. They will prioritize finding ways to do this.

Ideas to Supplement to Current Program Offerings

Traugott requested that the Committee discuss potential supplements to the Summer Program's current offerings (e.g., adding data science component to the curriculum). He explained that ICPSR Summer Program has started a conversation with SRC's Summer Institute about joint advertising. The Summer Institute enrollment has ticked up. Perhaps closer integration of ICPSR Summer Program and Summer Institute coursework in the future is possible. Stephanie Carpenter provided some additional detail about methods for cross-collaboration in events etc. Another opportunity for collaboration is the Blalock Lectures.

Summer Program Advisory Committee
Traugott transitioned to a discussion of the proposed Summer Program Advisory Committee. Traugott and Levenstein have discussed ways to leverage such a committee. The original plan was to have members of the Advisory Committee drop in at their convenience during the summer for a day or two to observe, and then convene a meeting in conjunction with the October Council Meeting. Advisory Committee members could collect observations and opinions to inform the next iteration of the summer program.

Status on Search for the Summer Program Director
Levenstein provided an update on the search for future leadership of the Summer Program. She has been having discussions with political scientists about what kind of position this should be. There are different ways of configuring the position, which would affect the job description and evaluation criteria used to assess candidates. Levenstein has received a lot of feedback on these questions.

Key points in the discussion of the Education Committee summary:
--ICPSR should consider adding more Teaching Assistants for online-only summer courses to ensure the best experience.
--ICPSR should consider researching and using a single third-party tool to support all online classes (e.g., Coursera), rather than having each instructor make decisions about the tools they'll use.
--ICPSR should ensure that data privacy training is provided as a part of the Summer Program, either in a course or at least in Blalock lectures.

Technology Committee

Council attendees: Bobray Bordelon (Chair), Jon Cawthorne, Lisa Cook, and Ken Smith

ICPSR Staff: Dharma Akmon, Trent Alexander, Jon Brode, Alina Conn, Abay Israel, Daphne Lin, Jared Lyle, Trisha Martinez, and Dan Pritts

Hiring
A new application architecture lead starts April 6, 2020. We are also hiring a senior devops engineer and aim for them to start before the end of April. CNS is planning to begin hiring to prepare for COA3D and will hire some each quarter to spread onboarding out and be prepared if we get the award.

Accomplishments
Sprint cadence has been changed to two weeks and have seen an improvement in quality of the work. CNS has also implemented estimation office hours, improving communication between those requesting the work and software developers. Our sprint planning has also improved, we have released several new websites, and the migration of websites to the new platform is going well.

Challenges
Big growth of staff will be needed in coming months for COA3D. CNS has about two-to-three times the work that they have resources to accomplish in a given sprint. They have dedicated attention to tracking and being transparent about tech costs.

Because of the high amount of work to resources ratio, CNS employs the following algorithm for decisions on what gets done in a given sprint: 1) deadlines; 2) egregious bugs; 3) number 1 tech priority: Curation tools.

CNS asked the committee for guidance on the following:

- Are there best practices that council knows about for getting things through the queue
- Can we analyze the backlog to see where there might be opportunity for funding?
- How to dynamically add staff as needs dictate—some use of contract staff. Onboarding is costly and we don't want to fill ongoing needs with short-term staff

StatSnap update was requested: development has been deprioritized for now. We want to do subsetting and weights and restricted online analysis. These ideas are part of COA3D, so would be a priority if awarded. There was general agreement that we should move Statsnap out of Beta now without further development work at this time.

Key points in the discussion of the Technology Committee summary:
--Council needs more information on how ICPSR is managing IT work requests, also known as "tickets." The ICPSR IT Director described how tickets are prioritized every two weeks before each "sprint" begins. Her presentation made clear that ICPSR acts on less than half of the tickets that are considered at each meeting. The not-acted-upon tickets were presented as gray bars in a visual display. Council asked whether these "gray" tickets were the same tickets from sprint-to-sprint, and why we did not formally decline the work. ICPSR will provide updated and more detailed metrics on the types of tickets that are acted upon immediately, are acted upon after several sprints, and are never acted upon.

Budget Session

Jon Cawthorne (Chair) introduced the new Council finance and policy committee and summarized their charge. He presented updated FY2020 fiscal situation and preliminary FY2021 budget. For FY2020 ICPSR is expected to break even. The preliminary FY 2021 budget has a 300K surplus while the Summer Program expects a small 77K shortfall. The presentation included a pie chart that summarized 2021 expenditures at ICPSR across major spending categories.

Discussion
The impact of COA3D if funded would mean more revenue, but the percentages in the pie chart mostly do not change.

Council thought that the pie charts were a helpful breakdown of how ICPSR spends its money across major categories.

A question was asked about the algorithm used to determine ISR taxes, ICPSR's share of those taxes, and university support to ISR and ICPSR. There is a negotiation with the university led by the ISR director. ISR currently covers part of the ICPSR director's salary. The university supports a fraction of the RCMD (Resource Center for Minority Data) director's salary, support for federal statistical agency data initiatives, and support for training of underrepresented students in the Summer Program.

Council asked for a breakdown of Federal versus private funding sources that would be helpful for allowing the Council to weigh in on the funding mix at ICPSR.

The projected Summer Program shortfall of 77K may be very different than expected; 77K is a lower bound of the shortfall; if the Summer Program revenue is lower because COVID-19 further reduces Summer Program participation. Council encouraged ICPSR to consider online instruction as a model for the Summer Program should COVID-19 continue to affect face-to-face instruction through the summer months.

Panel: Data anonymization in the 21st century

Lindsey Malcom-Piqueux (California Institute of Technology and ICPSR Council) chaired the panel discussion, "Data anonymization in the 21st century." Simson Garfinkel (U.S. Census Bureau) opened the panel (presenting remotely from Washington, D.C.) about de-identification versus data anonymization, the difference between working with structured data versus unstructured data, and putting differential privacy and de-identification into context.

"Anonymization is an aspirational description of what we want rather than a process of what we follow," Garfinkel noted. "You can describe what you do to the data, but you can't describe how the data results because that depends upon who receives the data -- what they can do with it."

Garfinkel described the motivations for re-identification, including testing the quality of de-identification, gaining publicity or professional standing, embarrassing or harming organizations, gaining direct benefits from the re-identified data, and causing embarrassment or harm. He noted that academics re-identifying datasets are a significant threat to de-identification and should not be underestimated since they have more time, more data, and more money than those performing the de-identification, as well as all the datasets of the future.

Garfinkel explained the challenges of de-identification by explaining that even if you remove names and obvious identifiers, just one variable left behind can trivially re-identify the dataset. He noted that you don't know what external dataset an attacker will have; information can be re-identified without obviously being an identifier. Garfinkel then described the few ways to future-proof datasets against de-identification: adding noise, publishing synthetic microdata, and not publishing microdata at all.

Looking forward, Garfinkel noted that non-structured data (e.g., free-formatted medical text, photos, medical imagery, GIS) will pose a big challenge for de-identification. All represent modalities where traditional identifiers may not be present, but the high dimensionality of

underlying data means that it can be re-identified. He noted that even though technologies have been developed to address re-identification risk for these modalities, knowledge about de-identification modality is not well-distributed. He gave an example of a 2019 journal article from the New England Journal of Medicine that "discovered" there is a risk of identifying people from MRI and CT scans, even though the risk had been an identified problem for more than ten years. Regarding text de-identification research, Garfinkel said many have taken the approach of finding and censoring identifiable data. One approach is to look for sensitive tokens and replace them with labels (e.g., "Patient Name"). Another approach is to replace with surrogates (e.g., "Gene Smith"). However, Garfinkel noted there is really no way to be sure that an attacker cannot re-identify someone, even if you remove the identifiers from unstructured text.

Garfinkel closed his presentation by noting that differential privacy is not a de-identification technology, and that it only makes sense for structured data. But what we know from the math of differential privacy, Garfinkel said, is that every column is potentially identifying, which is why every column has to be processed with noise and that goes against the de-identification ethos. He noted that another issue is that differential privacy has the concept of privacy loss versus utility or accuracy trade-off. Differential privacy allows you to choose that trade-off; de-identification doesn't allow you to choose that trade-off.

H. V. Jagadish (University of Michigan MIDAS and Computer Science) spoke next about anonymity in practice. "You really cannot have anonymity with individual records with microdata. It's just impossible," he warned. "The reason it's impossible is because there's enough other data...you can somehow join it with some external data and figure out what's going on." He gave the $1 million Netflix Prize as an example, in which "an in-the-closet lesbian mother sued Netflix for privacy invasion, alleging the movie-rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its" contest (quoted from https://www.wired.com/2010/03/netflix-cancels-contest/, which was not quoted by Jagadish in his presentation).

Regarding what we can and should do, Jagadish recommended having simple minimal barriers that avoid re-identification by accident and reduce the temptation to re-identify. He recommended sharing data through a responsible repository.

Jagadish closed his presentation by noting that differential privacy is the only technique that has provable privacy guarantees, but that it is complicated, it has to be done right, and it works only in very specific circumstances. He said that the Census, for example, is set up for exactly the conditions for which differential privacy works. But he also emphasized that he doesn't think that that differential privacy is a standard that should be applied for all of the other data sets that are collected or shared, including with ICPSR. Simpler forms of de-identification and relying on other researchers not re-identifying might be an appropriate way forward.

Rujuta Umarji (ICPSR) discussed the process ICPSR uses during curation to reduce the risk of reidentification. She noted that all curated studies receive a disclosure risk review, which considers both the access level for the data and the curation level. Lower curation levels may

just mask any responses that look potentially disclosive, while at the higher levels of creation there is more nuance to the remediation strategy. Access level determines whether data will be released as public or restricted.

Umarji described the disclosure review questions considered by ICPSR staff:

- What is the methodology that is being used?
- What is the population?
- What is the sample and is it a special or vulnerable sample?
- Are there subject specific particularities, like crime and health data?
- Are there special consent issues or issues of protecting respondent confidentiality?
- Are there linkages to previous studies or outside information?
- Is there sensitive information such as drug use, school records, or criminal records?

In addition to these bigger picture issues, Umarji noted that curators will try to identify both direct identifiers (e.g., names, birth dates, death dates, some geography, telephone numbers, addresses) and indirect identifiers (i.e., combinations of variables that can form a unique profile of an individual). For public use data, ICPSR typically masks all direct identifiers. For restricted use data, while direct identifiers are expected to be removed, ICPSR might leave the potentially disclosive indirect identifiers in the data since they will be disseminated via a secure download or via the Virtual or Physical data enclaves.

Umarji indicated that when remediating indirect identifiers, ICPSR may mask responses (especially if a study needs to be released quickly), as well as top- and bottom-code numeric variables if there are outliers. Additionally, ICPSR might collapse the variable into categories, or if there are already categories, collapse them further. When reviewing geography, if state is the smallest geographic unit, ICPSR considers if there are variables with fewer than 3 observations. For smaller geographical areas, ICPSR looks if they are variables with fewer than 10 observations. Population size is also assessed. Similar cell size rules are considered for other potential indirect identifiers, such as demographic variables (e.g., race, ethnicity) with fewer than 10 responses. Guidelines are also in place for reviewing income, physical characteristics (e.g., height, weight, BMI), medical history, household size, and sensitive behavior (e.g., drug use, age).

Umarji noted that when making decisions about remediation, ICPSR considers both re-identification and possible harm. For higher levels of curation, ICPSR also uses an additional worksheet that helps curators think through where variables might fall on these scales.

Yajuan Si (University of Michigan ISR/SRC Survey Methodology) presented "Confidentiality and privacy protection after record linkage: laying the groundwork for synthetic record linkage." She discussed synthetic data, synthetic geospatial measures, and synthetic record linkage. Regarding the synthetic geospatial measures, Si discussed a MiCDA pilot grant to develop and evaluate procedures to create survey data linked with synthetic geographic data designed to address confidentiality and privacy as well as analytic concerns. The project will extend existing synthetic data generation techniques using multiple imputation by linking the

Panel Study of Income Dynamics (PSID) to geographic data (counties, census tracts, blocks) from the U.S. Census; generating synthetic geospatial measures and generating synthetic PSID data.

After the panelist presentations, an audience member asked about the privacy implications of genealogists, who want to be re-identified. Panelists responded by noting the higher burden for organizations who collect private data and make it public. The more data that is publicly available, the less effective synthetic data is for protecting confidentiality.

Maggie Levenstein asked the panelists whether there are easy or different things ICPSR can do to improve its data disclosure review and remediation practices. An audience member directly followed Maggie's question by asking whether there is a risk to remediating disclosive data based on cell size. Noting that it does not give the guarantees of differential privacy, Jagadish responded by saying there is nothing wrong with using cell size, especially since it is a cheap and quick and easy first cut. He also encouraged the consideration of using the plain addition of noise.

Panel: Social & Legal Challenges of Restricted Data Access

Dave Armstrong, Western University, chaired and welcomed participants to the session.

Key Summaries from Panel Participants

David Grimm, UM Office of General Council, addressed contractual issues by first noting that as a data organization, you don't want to get to the point where you are talking about the contract specifications (because there has been violation); you must as an organization include and execute education and process components as part of the legal agreement.

He defined three components of contractual and legal issues:

- Contract: should be written in plain language versus a lot of legalese; should have appropriate (specific) language for the situation versus borrowed and broad language from other related or non-related contracts; needs to be enforceable and organization must be willing to enforce the penalties
- Education: ensure that those signing the document know what they are agreeing to (plain language from above) and what to expect; set expectations early and often so those signing the contract fully understand what it says/what they are agreeing to and the penalty process.
- Process: get the right person(s) to read the document and be on the hook for it since often those that are signing are different from those that are accessing the data; obtain the authorized organizational signature and also a signature(s) from the data user(s) so they actually comprehend what the contract says.

Tim Mulcahy, Westat, covered social, legal, and technological challenges of restricted data access. He postulated that the "pendulum has swung from a focus on data security to "extreme" researcher convenience." He noted that researchers today seem to have a lack of appreciation for security restrictions and want it now and for free. There is disregard for safe use thus requiring

data organizations to set clear guidelines for safe use (including education on professional data stewardship). There is a need to outline (educate) the rules of fair game for virtual access to secure data in what he called the data "Wild West" environment.

Mulcahy noted several issues to address in this environment:

- Ego - a sense of researcher entitlement and the current convenience mentality - Organizations must answer (educate) questions including: What's the big deal (with respect to data security)? Why can't I just have it now (immediately) if I just sign your agreement?
- Legal challenges: policies and sanctions need to be carefully constructed to ensure they are working and will continue to work; researchers must read the agreement; how can they prove to you that they have read the agreement?
- Training regimen - is it working? Does it go beyond checking the box? Does it establish trust that the data user understands?
- Technological: As technology grows, are we keeping up? The cloud model has produced challenges that we must adapt to.

How might data organizations enforce rules in the "Wild West?"

- Revisit idea of reducing access like the FSRDC does
- Sanctions must be implemented, and it must be made known that sanctions have been implemented (the data organization takes rule violations seriously)
- Training must be not just electronic but must have testing and humans involved
- Researchers must know/see we are monitoring and tracking; big brother is indeed watching.

Barbara Downs, US Census Bureau, addressed: Challenges & Failures of What We Do. She began with an overview of the Federal Statistical Research Data Centers (FSRDC) and noted the environment is collaborative but secure within the FSRDC technical environment. As a means to consider for our own data organizations, she highlighted and explained the FSRDC goals: Safe project; safe people; safe settings; safe data; safe outputs. She focused on safe people and settings for this discussion describing the elements associated with each:

- Safe People
  - Background investigation of applicants in proposal (project team); extensive review; rigorous and not all pass
  - Must take confidentiality for life (forever); sanctions will apply (and indeed, sanctions have taken place)
  - Data use agreement: data users are responsible for their own behavior in understanding and following rules and having sanctions imposed
  - Research orientation: there is a quiz that must be passed; otherwise no access is given
  - Annual data stewardship training is required; it is not enjoyed by researchers, but continued access is conditional on completion
- Safe Settings (Physical)
  - Access controls (badge, etc.)
  - Monitoring - video monitoring always in use
  - Internet - no access
  - Materials control - no materials leave facility without review
  - Printing - strictly controlled

She noted that sanctions do apply and are enforced. Researchers who violate rules are suspended and likely will need to complete entire retraining or may require an in-person monitor to be reinstated. Expulsion has been used.

Trent Alexander, ICPSR, added that ICPSR has hired a privacy officer to concentrate on the issues covered since the demand for restricted-use data continues to grow. One of the first steps taken was to redo and require that new training be completed.

Additional Points from Discussion

Discussion on sanctions

- Census can impose sanctions upon the data user that violates agreements; other entities don't have such leverage, so UM's strategy is to get the data user's institution (versus solely the data user) included in the liability to emphasize the seriousness of sanctions (some peer and institutional pressure).
- What about training students on these issues? The sense is that it is not well-codified for training student researchers.
    - A recent research project found that syllabi across a large number of classes/institutions covered no data stewardship/data security instruction nor their (the student's) responsibility; there was no emphasis on protecting research subjects and thus no sense of self-responsibility would likely be assumed.
    - Data organizations should be sharing "real" instances of bad behavior and the sanctions that were applied to instruct on the seriousness and to establish the culture of responsibility. The idea is to use actual cases to educate students and researchers on data security and responsibility.
- GDPR status on research data - generally, research is always a compatible purpose; there is some case law now that is supporting use of data even if consent is problematic, because research is "always" a compatible purpose (but you still have to notify the research subjects - which is problematic). If US data, not much to worry about in terms of research data - can go forward and share.

Overall recommendations regarding ICPSR's focus on privacy and restricted data:

Our conversations about privacy have largely focused on government data and data produced by academic institutions. At a future Council meeting, we should focus on the privacy approaches of data produced and managed in the private sector (e.g., Ancestry.com, Facebook, other commercial data).