

## Council Meeting Minutes

June 7, 2018

### Attendees:

**Council Members:** Bobray Bordelon, Lisa Cook, Jane Fry, Michael Jones-Correa (Chair), Chandra L. Muller, Verna Keith, Robert Stine, and Katherine Wallman

**ICPSR Staff:** Dharma Akmon, Trent Alexander, JD Alford, Rita Bantom, Shuming Bao, Johanna Bleckman, Dieter Burrell, Stephanie Carpenter, Alina Conn, Edward Czilli, Linda Detterman, Libby Hemphill, Lynette Hoelter, Dory Knight-Ingram, Sanda Ionescu, Abay Israel, Susan Jekielek, Kilsang Kim, Kathryn Lavender, Lauren Lee, Susan Leonard, Maggie Levenstein, Jared Lyle, John Marcotte, Elizabeth Moss, Tom Murphy, Justin Noble, Asmat Noori, Michelle Overholser, Amy Pienta, Jukka Savolainen, Saundra Schneider, Michael Shove, Fillippo Stargell, David Thomas, Vanessa Unkeless-Perez, and Diane Winter

### Approval of Minutes

ICPSR Council Chair call for the minutes of the March 2018 meetings. Council voted unanimously to approve.

### Director's update

#### a. Archonnex

Our number one technology priority is to complete the move of curation to Archonnex. The move is part of our Strategic Plan. Staff in curation, metadata, and I&T are engaged in developing systems for our future needs.

#### b. Acquisitions

ICPSR's acquisition goals remain focused on core social science areas that will have a broad impact on future research. Staff has initiated several new acquisitions strategies:

- Tracking depositor data in new software to facilitate follow up.
- Use OpenICPSR for replication data and to increase name recognition within the research community, but not for general deposits.

#### c. Sponsored Projects

The National Archive of Criminal Justice Data, which ICPSR has hosted for 30 years, is expected to be competitively bid in the next quarter. We will be submitting a new proposal to the Mellon Foundation to support a new data collection project to measure the impact of liberal arts educational experience.

ICPSR and the State of Michigan are in the initial stages of discussions to establish an archive of state administrative data. The archive will be used by researchers working with state agencies to facilitate access to interagency data.

Several initiatives have been implemented to improve our proposal process:

- A new form has been created for proposal vetting to ensure our proposal writing and research administration resources are being used strategically and alignment with the organization's goals.
- Administration has been improving the budgeting process to allow for better tracking of resource allocation.
- Improvements to the quality of template texts for proposals.
- The creation of a proposal review group to edit proposals prior to submission.

#### d. Organization

ICPSR has filled many vacant positions, including computing, communications, curation, and managers. Currently, hiring has been suspended, with the exception of Summer Program temporary staff, until more information is available regarding the fiscal year 2019 budget and sponsored funding.

The reorganization of curation and project management units has allowed for more flexibility between projects, giving staff more security. The leads of each Functional Unit (FUnC) meet weekly to facilitate interunit communication. A smaller strategy group meets less frequently to address long term issues of the organization. (Alexander, Detterman, Levenstein, Murphy, and Pienta currently constitute the Strategy Group).

### **Update on Strategic Plan**

ICPSR Director and Strategic Plan Consultant have been meeting with working groups to discuss progress and identify problems.

## **Overview of financial situation**

An updated income statement and proposed budget were distributed. The updated versions adjusted for capital equity depreciation and had no change in spending plans.

The proposed budget for fiscal year 2019 assumes a constant level of membership activities. The projected decline in sponsored funds is a conservative estimate. It does assume that the NACJD and Education archives will have their funding renewed. Any other pending proposals are not included. The budget also includes a \$200,000 increase in ICPSR's Provost Tax; implementation of the tax will be confirmed in July 2018.

We will be recommending that Council approve no increase in membership rates for 2020, as new classifications and pricing went into effect this year. In the future we expect to increase rates consistently at 3% to reflect increases in expenses.

## **Education Committee**

**Council:** Bobray Bordelon, Jane Fry, Verna Keith, and Katherine Wallman

**Staff:** Dieter Burrell, Scott Campbell, Stephanie Carpenter, Edward Czilli, Linda Detterman, Libby Hemphill, Lynette Hoelter, Maggie Levenstein, Sandy Schneider, and Fillippo Stargell

## **Education Strategic Plan**

Maggie Levenstein opened the meeting by saying there is no formal agenda. She discussed plans for the Education Strategic Plan. She also introduced Libby Hemphill as the leader of that strategic planning process. Maggie indicated that she would like the education strategic plan to integrate ICPSR Summer Program more closely with ICPSR's other educational and training plans.

Libby stated that the strategic planning process was still in its beginning stages, with the idea to first understand where education occurs in ICPSR. A rough spreadsheet was sent to different groups at ICPSR to find out what kind of training is happening and when it happens to document the scope of what ICPSR is already doing and to identify potential areas for closer coordination.

Maggie gave a brief overview of the new ICPSR structure to give some context to the different groups on the spreadsheet. She believes the Summer Program is a good way to test new training ideas and strategies before offering them to the ICPSR membership more broadly. Maggie said the Summer Program is trying to incorporate more ICPSR data into their classes to expose participants to ICPSR.

Libby brought up Jupiter Notebook, a system that runs in a web browser. Jupiter Notebook allows users to compile and run code, as well as analysis of the code, for easier sharing among colleagues or students. She highlighted its use in teaching and reproducibility of research, and said the data community is starting to embrace it. Also emphasized the importance of ICPSR as

leader; ICPSR's use of Jupiter Notebook would help to educate the research community about its advantages.

Linda Detterman asked Libby if an audience analysis was in the strategic planning process. Libby noted that ICPSR targets multiple audiences including students and Principal Investigators and other groups. Libby said it is important for ICPSR to decide if it wants to serve everyone or build the audience it already has. Verna Keith said if ICPSR is targeting undergraduate students then it is also important that ICPSR targets the people who teach undergraduate courses. Libby agreed, saying that how ICPSR reaches groups is important as well when it comes to resources.

Maggie discussed an upcoming meeting on improving empirical research skills for undergraduates. Between a quarter and a third of ICPSR downloads come from undergraduates. Recognizing undergraduates and undergraduate instructors as central to ICPSR educational activities requires a shift in thinking.

Linda brought up the Data Driven Learning Guides as an excellent way to introduce people to ICPSR data.

### **Outreach and New Audiences**

A discussion was held on the topic of statistics literacy, with Bobray Bordelon saying some institutions require students to take courses on the topic, and there is a need for statistical literacy in the general public. Libby reiterated that need, and Maggie said the State of Michigan is interested in sending its staff to short workshops to strengthen their statistical and data management skills.

Lynette Hoelter gave an update on StatSnap. The goal of StatSnap is to offer a sustainable application that will replace the functionality currently provided by the SDA on the ICPSR site. The initial goal is crosstabs and simple descriptive statistics. Additional functionality will be added later.

Maggie asked if the Summer Program should target particular groups and if so, which ones: undergraduates, students just entering graduate school, or international students.

### **The Summer Program and Housing**

Sandy was asked about the Summer Program's diversity efforts. She said the Diversity, Equity, and Inclusiveness initiative holds great promise in addressing underrepresented groups. She also said participants come to the Summer Program in part to learn techniques and methods, but also to learn how to teach, as a number of participants go on to teach undergraduate statistics or basic methods. She mentioned the benefits of teaching policy makers as well in order to reach the general public more easily. Reputation and stature of the Summer Program is important. The opportunity to reach these other audiences is there.

Bobray thought a sponsored short workshop on data journalism would be popular, and the room agreed.

Libby and Maggie both talked about the benefit of new graduate students coming to the Summer Program and subsequently entering their first year of graduate school with stronger quantitative skills.

Verna asked how the Summer Program rates were calculated, which led to a discussion on housing costs in Ann Arbor. Sandy stated there is not a lot of movement on dorm space from the University, but that the new Summer Program survey would ask participants if they would be willing to spend money to stay in a U-M dorm. There is also possible space at Eastern Michigan University, that transportation would be an issue.

### **Digital Communications**

Bobray talked about the benefit of finding ICPSR webinars on YouTube, which led to a discussion of outreach strategies. Linda said email is still very useful for what ICPSR does, and that social media can be used in ways email cannot, such as repetition of announcements or content. Jane Fry said linking back to datasets talked about in videos could be very useful, either in the description or through annotations.

Maggie ended the meeting by asking Council to think not just about what new things the Summer Program should do, but also what it might stop doing.

### **Acquisition Strategies and Metadata Priorities**

**Council Members:** Bobray Bordelon, Michael Correa-Jones, Jane Fry, Verna Keith, Chandra Muller, Robert Stine, Katherine Wallman

**Staff:** Trent Alexander, Sarah Donetti, Libby Hemphill, Sanda Ionescu, Margaret Levenstein, Jared Lyle, Elizabeth Moss, Amy Pienta, Jukka Savolainen, Annalee Shelton, Vanessa Unkeless-Perez

### **Acquisition Strategies**

Amy Pienta, Director of the Business and Collection Development (BCD) Unit, provided an overview of the unit's accomplishments, challenges and goals. Since March 2018, there was an increase in the rate of deposits to the General Archive, as well as to openICPSR. Staff have started using ProsperWorks, an online system for managing relationships with customers which should facilitate deposits. Staff are working on establishing a more coherent acquisitions relationship between ICPSR and openICPSR, and are looking at ways in which openICPSR may help ICPSR establish a presence in different communities. Pienta reported that staff are trying to find new audiences and obtain data that are more diverse, but also of good quality and suitable for reuse.

Council asked if the deposit process is still too onerous and if some of the data creators want to keep track of how and when their data are re-used. Maggie Levenstein noted that the deposit form has been improved, but it is still not easy to use, and is under revision. Pienta noted that ICPSR staff is available to help with deposits, and that ICPSR is re-crafting the restricted data deposit agreement and trying to expedite the review time. Pienta also noted that ICPSR is developing and refining the process by which we interact with the data creator, including providing them with updates regarding their data usage.

Libby Hemphill mentioned that ICPSR has been contacting commercial entities about getting their “older” data (not in current use) and making it available. These entities include Twitch and Twitter; we plan to talk to Facebook.

Council noted that ICPSR could also target government departments that are closing down and rescue their data. Levenstein noted that ICPSR received funding to support at-risk data identification and acquisition through DataLumos.

### **Metadata Priorities**

Jared Lyle, Director of the Metadata and Preservation Unit, provided an overview of how ICPSR provides information-rich metadata. He then asked the Council how they would like to see data described in order to find and understand data collections, if there are other places ICPSR can learn from who do a good job describing and surfacing collections, and what ICPSR can do better to help others describe their collections prior to archiving them.

Council noted that ICPSR metadata is of high quality and often cited as a standard within the data and library communities. They also noted that some other repositories and data providers are providing quality metadata, but not to the extent ICPSR does. One example was a provider coding data to the variable level, but with missing metadata about sample size, geographic coverage, or years covered.

Council indicated some of the most useful metadata fields when using ICPSR collections include: producer, abstract (since title is not enough to accurately represent the data content), years covered, sample composition and size, geographic coverage (i.e., which part of the country), if the study is part of a series, and main variables.

Council encouraged ICPSR to do more with series records. Two examples were provided. The first was to include series records in weekly emails of new and updated data collections. Currently, only individual collections are included. The second example was to include useful openICPSR deposits (e.g., Uniform Crime Reporting (UCR) data collections) in ICPSR series.

Council encouraged ICPSR to use the summary field in metadata to pay more attention to capturing the topics covered in the data, especially when there are collections about unusual or unexpected topics, like the questions on religious issues in AddHealth or the ANES series.

Levenstein asked the Council for their thoughts about ICPSR crowdsourcing for metadata enhancement.

Council noted that it is important to determine how the external contributions are vetted, and that ICPSR might want to first experiment with crowdsourcing with a trusted community of experienced users.

Levenstein asked the Council for their thoughts about ICPSR's role in leading the DDI Alliance. Council noted that ICPSR's support of the DDI specification is important for the data community, and a strategic asset for ICPSR. They emphasized that it is important for the organization to continue to play a leadership role in the DDI community. Council indicated that DDI adds value to ICPSR metadata and that many of the user-oriented features on the ICPSR website are made possible by the use of DDI.

Council noted that putting metadata in place is now a high priority within the federal statistical systems of the United States. ICPSR's metadata, including DDI, is a good way to demonstrate its comparative advantage.

## **Technology Committee Meeting**

**Council Members:** Michael Jones-Correa, Chandra Muller, and Robert Stine

**Staff:** Dharma Akmon, Trent Alexander, Johanna Bleckman, Wendi Fornoff, Sanda Ionescu, Abay Israel, Kil-Sang Kim, Maggie Levenstein, Jared Lyle, John Marcotte, Elizabeth Moss, Tom Murphy, Asmat Noori, Amy Pienta, Matthew Richardson, Anna Shelton, Allison Tyler, Harsha Ummerpillai, and Vanessa Unkeless-Perez

### **Curation**

Tom Murphy opened the meeting by emphasizing that our top technical priority is the movement of curation work into Archonnex. Curation includes data manipulation, clearing documentation markup, and "turnover" of data into the dissemination system. Current work is "administration" parts of curation, where data is brought into the curation area and sent out to the dissemination area/website. That work will be completed in 2018. In 2019, curation improvements will focus on individual file/project manipulation.

Abay Israel was introduced as our project manager/agile coordinator. Abay coordinates the work of a group of product owners, which are ICPSR subject matter experts who spend part of their time supporting IT work. They work with stakeholders and submit "tickets" in JIRA software. Abay works with IT staff to transform these into technically-oriented tickets and structure those tickets into "sprints" of approximately 1 month. Abay demonstrated the software that shows which work is happening in each sprint and took a detailed look at a single ticket.

Council asked about efforts to automate curation work and about efforts to bring in new types of data. The 2019 curation work will automate some tasks, though the initial focus of the work will be to rationalize the work and processes across archives and update any processes that are based on one-off scripts or otherwise not sustainable.

Abay showed a "curation at ten thousand feet" process diagram. This is a visual representation of what we're planning for this year and our priorities. He showed a slide that explains the five phases of the curation system: administration, planning, work, quality check, release. All but the "work" phase will be done this year. Within "work" only the metadata editor will be done this year. Finally, Abay also showed a chart that explains the reasoning behind how we prioritize tickets. It clarifies what components have the highest priority. The data editor/document editor is unlikely to be prioritized this year.

Council asked the larger group how they perceived this change and whether they were concerned. One curator pointed out that there are a lot of repetitive tasks in curation. In a 30-part study where you find a typo in an investigator's name, you get to make that change 30+ times. A human gets to spend a half-day doing something that would take a machine 5 minutes. He said he looks forward to more focus on things like disclosure analysis, creation of public use datasets, etc. that have curators use their unique skills, rather than lower-level work. Other curators have expressed concerns about job security. Trent said that it is not a goal to reduce the size of the curation staff but rather to curate more data more quickly.

ICPSR staff asked about the "backlog" of membership curation. Trent answered that there had been a backlog of 65 studies. They were reviewed and only 15 were curatable; those were curated. The rest are preservation only. Under the new process with multiple curator levels every single dataset that comes in will be processed and released within two weeks. This is the "beta" level of release. These datasets will be available for more intensive curation if warranted.

Council suggested marketing any automatable routines we develop to others. More generally, they suggested that we highlight our standards and accomplishments better. We are underselling ourselves and what makes us unique in an increasingly crowded marketplace.

One ICPSR archive director noted that federal funders are often focused on publications, and don't see much value in curation. Different audiences have different views on curation. Funders focus on things like learning guides, bibliography, and add-ons. They don't want their data touched/edited and such. Rather than fixing data, we should report errors...have an errata system. Council agreed that this was a good idea, but emphasized that we may need to look beyond NIH and federal funders.

## **Credentialing System**

ICPSR staff described our Credentialing System project, "Researcher Passport," which is funded by the Sloan foundation and led by Johanna Bleckman and Allison Tyler. A researcher applies for a passport which has identifying information, CV, affiliation, etc. Much of this info will be verified. The researcher can then apply for data providing the digital passport to the data provider. The data provider issues a "visa", which grants access to the data. Allison pointed out that this is a huge benefit to researchers who repeatedly access data from the same repository/institution. Verified identity saves a lot of time for both parties. The passport should be available in Fall 2018.



Council expressed a concern about data misuse. Would the system record improprieties? Johanna said there will be a record of misuses. The system will accept allegations and there will be a mechanism for judging and attaching such things to the passport. Council asked how this is different from the UK system. Allison affirms that it's similar to the UKDA. Allison surveyed multiple international institutions. Council expressed concerns about legal issues. Johanna suggests that U-M general counsel will review them. Council asked about how long offenses would remain on record.

One Council member compared this to school transcripts. Suggested that we should keep credentialing and violations as an entirely separate component. NACJD archive director said that he expects the credentialing system will remove a lot of the burdens for sensitive data application.

### **StatSnap & DLRep/Linkage Library**

StatSnap is an online data analysis tool under development. Led by Lynette Hoelter, it is expected to be released by the end of 2018.

Linkage Library is a repository to facilitate the sharing of code, techniques, and data for the record linkage research community. It will be released by the end of 2018. ICPSR hopes to re-use the commenting/code-sharing functionality on other sites.

Council requested a plan to market Archonnex beyond ICPSR. All agreed that this would be discussed at the next Council meeting.