

ICPSR **M** | ISR BULLETIN

4 Fun in the sun

Splash into summer with ICPSR's Summer Program in Quantitative Methods of Social Research



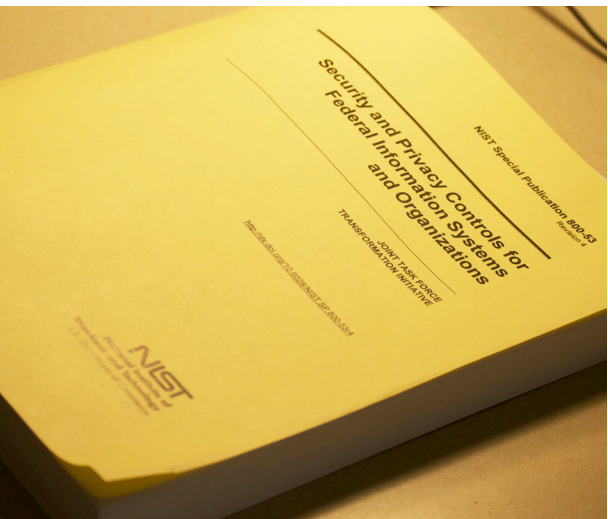
Archiving in Africa

ICPSR teams up with DataFirst to train data archive staff at University of Cape Coast in Ghana



10

ARCHIVE CLEARS RIGOROUS FISMA COMPLIANCE PROCESS



“We’re managing confidential data for sponsors and we’re committed to making sure the tools and systems we’re using are secure. This is proof that we’re living up to our responsibilities and doing it the right way.”

George Alter, ICPSR Director

The [National Addiction & HIV Data Archive Program \(NAHDAP\)](#) recently cleared a rigorous Federal Information Security Management Act (FISMA) compliance process with flying colors.

The National Institute on Drug Abuse (NIDA) issued an authorization to operate, or ATO, for the NAHDAP on February 12.

Efforts were led by Asmat Noori, assistant director of Computing and Network Services, and John Marcotte, an ICPSR archive director. As he flipped through the 1-inch-thick FISMA manual titled *Security and Privacy Controls for Federal Information Systems and Organizations*, Noori said, “all together from the time we started until we got our ATO from NIDA would be about a year. But once we got into it, we realized that 90 percent of requirements we already met. A lot of the work was creating proper documentation and procedures. We had about 700 standards we had to consider.”

CNS Director Tom Murphy said this achievement is significant for both ICPSR and the university community at large, and he believes ICPSR’s experience and guidance can help others meet FISMA standards. “FISMA requirements are really becoming more prevalent in everything we do,” Murphy said.

Murphy and Noori noted that FISMA compliance is not as simple as following a list of guidelines and has lots of room for interpretation. “Different agencies may have different interpretations of the controls,” Noori said. Murphy said ICPSR will keep an eye on FISMA guidelines as they evolve.

Story and photo by Dory Knight-Ingram



Update: Five things to know

- 5

openICPSR is free to ICPSR members!

To meet a growing demand for open-access data management in the social and behavioral sciences, ICPSR launched [openICPSR](#) in early 2014. The service moved out of beta in February 2015.
- 4

Custom repositories for institutions and journals

openICPSR has been expanded to include openICPSR for Institutions and Journals. “The openICPSR team met with institutional representatives and learned that they wanted something they could brand, allowing them to showcase their research data, fulfill grant requirements and be able to pass an audit that could come from an entity such as NSF,” said Linda Detterman, ICPSR’s Marketing and Membership director. “We’ve set up a relatively low-cost way to do that in a way that requires little administrative work. The repository manager does not have to play gatekeeper on new submissions,” said Web Project Manager Matthew Richardson.
- 3

A recommended repository

The open-access journal *Scientific Data* recently added openICPSR to a list of recommended social science repositories.
- 2

openICPSR can handle private and confidential data

ICPSR will manage data sharing for data that must be restricted to protect the privacy of subjects. Restricted data in openICPSR will be available through ICPSR’s secure “virtual data enclave” under a data use agreement. Users will pay a small fee to cover the incremental costs of providing access to confidential data.
- 1

Enhance the value of your data

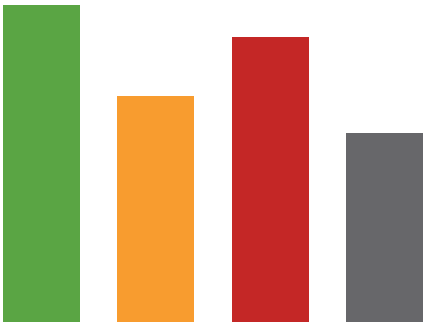
“If you choose, we can curate your data,” said Director of Curation Services Jared Lyle, “with checks on data quality, detailed codebooks, and translation into multiple statistical packages, which maximizes access and ensures long-term preservation.”

fun in the sun



As we enter another summer ripe with opportunities, hundreds are planning to soak up some knowledge at ICPSR's Summer Program in Quantitative Methods of Social Research. In the pages that follow is a wealth of information about opportunities to get ahead in graduate school, to overcome math anxiety, and to splash into some real data analysis!

Summer Program | Continued on page 6



SUMMER PROGRAM
AT A GLANCE

84
Courses
overall

98
Instructors

New statistical workshops in 2015

The following workshops take place in Ann Arbor, Michigan, unless noted otherwise.

Text Analytics, August 3–6, explores the application of text analytics methods across a range of substantive topics and research areas. Techniques include using raw counts of words, forming principal components from these counts, and building regressors from counts of adjacent words. The workshop will also explore proposed hierarchical generating models often associated with nonparametric Bayesian analysis. Because regressors derived from text may be difficult to interpret, the workshop will also show how to develop interpretive hooks from quantitative features.

Regression Discontinuity Designs, June 15–17, explores these useful and potent research designs for evaluating program impacts, policy issues, and societal problems. This workshop covers the basic principles, estimation, and interpretation of regression discontinuity designs, as well as their applicability across a broad array of substantive areas.

Designing and Conducting Experiments in the Laboratory, June 22–26, introduces participants to basic research design considerations, the problems of inference for incomplete designs (particularly field experiments), and canonical games common in experimental economics. To learn about experiments, there is no substitute for doing. Therefore, participants will work in four- to six-person

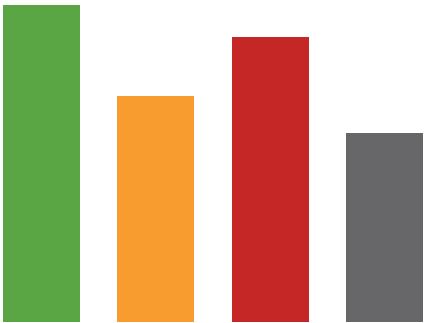
groups to design experiments that will be run during the workshop.

Qualitative Research Methods, August 10–12 in Chapel Hill, NC, presents strategies for analyzing qualitative data in the social, behavioral, and health sciences, as well as in the field of business. Both descriptive and interpretive qualitative studies will be discussed, as will more defined qualitative approaches such as grounded theory, narrative analysis, and case study. The workshop will consider how researchers develop codes and integrate memo writing into a larger analytic process.

Handling Missing Data Using Multiple Imputation in Stata, June 1–3, introduces multiple imputation (MI)—a simulation-based technique for handling missing data—and focuses on how to perform MI in Stata using the mi command. The three stages of MI (imputation, complete-data analysis, and pooling) will be discussed in detail with accompanying Stata examples.

Survival Analysis, Event History Modeling, and Duration Analysis, June 22–24 in Berkeley, CA, introduces participants to an applied methodology for the analysis and modeling of time-to-event data, which ensures proper handling of censored observations (not having experienced the event of main interest by the end of the study, for various reasons). Such observations, along with complete observations, arise frequently in the medical, social, and behavioral sciences when time elapsed until event occurrence is to be modeled and explained in terms of independent variables and predictors.





Sponsored workshops

Exploratory Data Mining via SEARCH Strategies, June 8–12, provides an overview of current techniques in exploratory data mining for quantitative research in the social and behavioral sciences. Exploratory data mining uses computational methods on large amounts of data in order to construct predictive models of behavior, in contrast to the standard hypothesis testing of many standard statistical techniques.

Secondary Data Analysis and the National Addiction & HIV Data Archive Program (NAHDAP), June 22–24, features brief presentations focused on the nuts and bolts of secondary data analysis followed by hands-on workshop sessions that include one-on-one consultation for conducting your own research.

Digging into the NSECE: Exploiting the Potential of the Household and Provider Data from the National Survey of Early Care and Education (NSECE), July 20–23, explores advanced topics related to the National Survey of Early Care and Education (NSECE), the first study of its kind in over 20 years. The NSECE provides a national picture of families’ non parental care utilization, as well as characteristics of both home-based and center-based providers for children birth through age 13. The NSECE

will help deepen the understanding of the extent to which families’ needs and preferences coordinate with providers’ offerings and constraints. The workshop is free for accepted participants, but space is limited.

Curating and Managing Research Data for Re-use, July 27–31, will instruct participants in best practices and tools for data curation, from selecting and preparing data for archiving to optimizing and promoting data for reuse. ICPSR social science quantitative datasets and UK Data Archive qualitative and cross-disciplinary data collections will serve as case studies, and participants will track the datasets as they make their way through the data assessment, review, processing, and curation pipeline. Emphasis will be placed on hands-on exercises demonstrating curation practices and on discussion for sharing local experiences and learning from others. Additional context and expertise will be provided through invited keynote lectures by research data experts.

Immigration, Immigrants and Health Conditions, Health Status, and Policies: Examining Multilevel and Multidimensional Models and Approaches, August 3–7, is designed for researchers interested in examining health disparities, conditions, and health equity as it pertains to immigrant populations.



4-WEEK SESSIONS:

39

courses

11

lectures

28

statistical workshops

SHORT WORKSHOPS:

45

courses

23

statistical and substantive workshops in Ann Arbor

22

statistical and substantive workshops in other locations

WORKING TO INCREASE DATA ACCESS IN AFRICA

Staff at the University of Cape Coast Ghana recently welcomed ICPSR Director George Alter, ICPSR data curator Doctor Ashe, and DataFirst Manager Lynn Woolfrey, who visited UCC in February to assist Sam Annim, Director of the Centre for Data Archiving, Management, Analysis and Advocacy (C-DAMAA), in the training of new archive staff.

One outcome of the trip is a new study, [“Social Learning, Social Influence, and Fertility Control \[Ghana\] \(ICPSR 35466\)”](#) released in mid May. Ashe was able to take C-DAMAA staff through each step of ICPSR’s data processing pipeline (see related story, page 24). Also, “on this trip, we had a visit with the national statistician of Ghana to talk about how C-DAMAA can help make data available to the academic community,” Alter said.

Both DataFirst and C-DAMAA would like to increase the academic community’s access to data across Africa. C-DAMAA is a microdata service center in West Africa that is housed in UCC’s Department of Economics. A grant from the University of Michigan’s African Social Research Initiative (ASRI) in 2014 has helped C-DAMAA and ICPSR staff visit each other’s institution in efforts to help C-DAMAA build up and maintain its repository.

Access | Continued on page 12

“It is wonderful to have an organization like the ICPSR to share their depth of knowledge and experience for a future African Consortium. I know from DataFirst’s history how this could benefit African researchers.”

Lynn Woolfrey, DataFirst Manager

DataFirst at the University of Cape Town provides African survey and microdata services to researchers and policy analysts. In 2008, DataFirst partnered with the Organisation for Economic Cooperation and Development to work with the Accelerated Data Program, which supports government data producers in developing countries in the production and use of national data. This work includes providing technical assistance, data management tools and training staff to curate data to international standards.

“The Ghana work we did was not initially part of our DataFirst/ADP partnership, but a project of George’s,” Woolfrey said. “We gave technical assistance to the University of Cape Coast to establish a data service to support quantitative research at their University. They use the same web-based data platform we use here to share microdata with researchers. This is the same software we promote and install at African government data producers — it is free and easy to use and build on.”

Woolfrey said creating an African consortium for sharing research data could leverage the skills at places like Cape Coast and the University of Cape Town to preserve and share African data. She will be co-facilitating (with ICPSR Director of Curation Services Jared Lyle) an ADP-funded

workshop in Ghana in June on establishing such a consortium. “It is wonderful to have an organization like the ICPSR to share their depth of knowledge and experience for a future African Consortium,” Woolfrey said. “I know from DataFirst’s history how this could benefit African researchers.”

The partnership between C-DAMAA, DataFirst and ICPSR grew out of the University of Michigan’s African Social Research Initiative. A group at ICPSR is in the early stages of creating a new guide to building a data repository so that others understand what it takes to build and manage their own repositories.

Story by Dory Knight-Ingram



TALES FROM THE FIELD

ICPSR’s mission is to advance and expand social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations. ICPSR staff often visit institutions and conferences around the world, helping to facilitate the sharing and preservation of data and to make the research community aware of the data curation and preservation resources available to them. Here is one of those “tales from the field.”

I was elated when I received the invitation to visit Ghana. I was honored to travel on behalf of ICPSR to assist in training at Ghana’s first data repository center. My main goal while at UCC was to guide C-DAMAA through ICPSR’s processing pipeline, processing the data themselves from start to finish. I also shared knowledge about ICPSR’s standards and processes and talked about best practices for data curation.

We were able to access ICPSR’s Secure Data Enclave (SDE) and use all internal tools successfully. By the end of my visit, the C-DAMAA staff were very familiar with the ICPSR processing pipeline. They worked on one study and were able to submit it for the first round of quality review. The study, [Social Learning, Social Influence, and Fertility Control \[Ghana\] \(ICPSR 35466\)](#), was released publicly shortly after I returned to Ann Arbor. In addition, I was able to show the staff how to create Survey Documentation and Analysis (SDA) on a previously released UCC study ([ICPSR 35296](#)) and how to use it to analyze data collections.

As expected with any major project, a few challenges occurred; however, ICPSR and C-DAMAA powered through technical glitches and were able to accomplish the overall goals. I am very happy that we were able to accomplish, which included introducing ICPSR to graduate students and faculty in the Economics Department, meeting with Ghana’s national statistical service, and consuming fine Ghanaian cuisine! Moving forward, I plan to remain in touch with the C-DAMAA staff and will act a resource should that need any additional guidance.

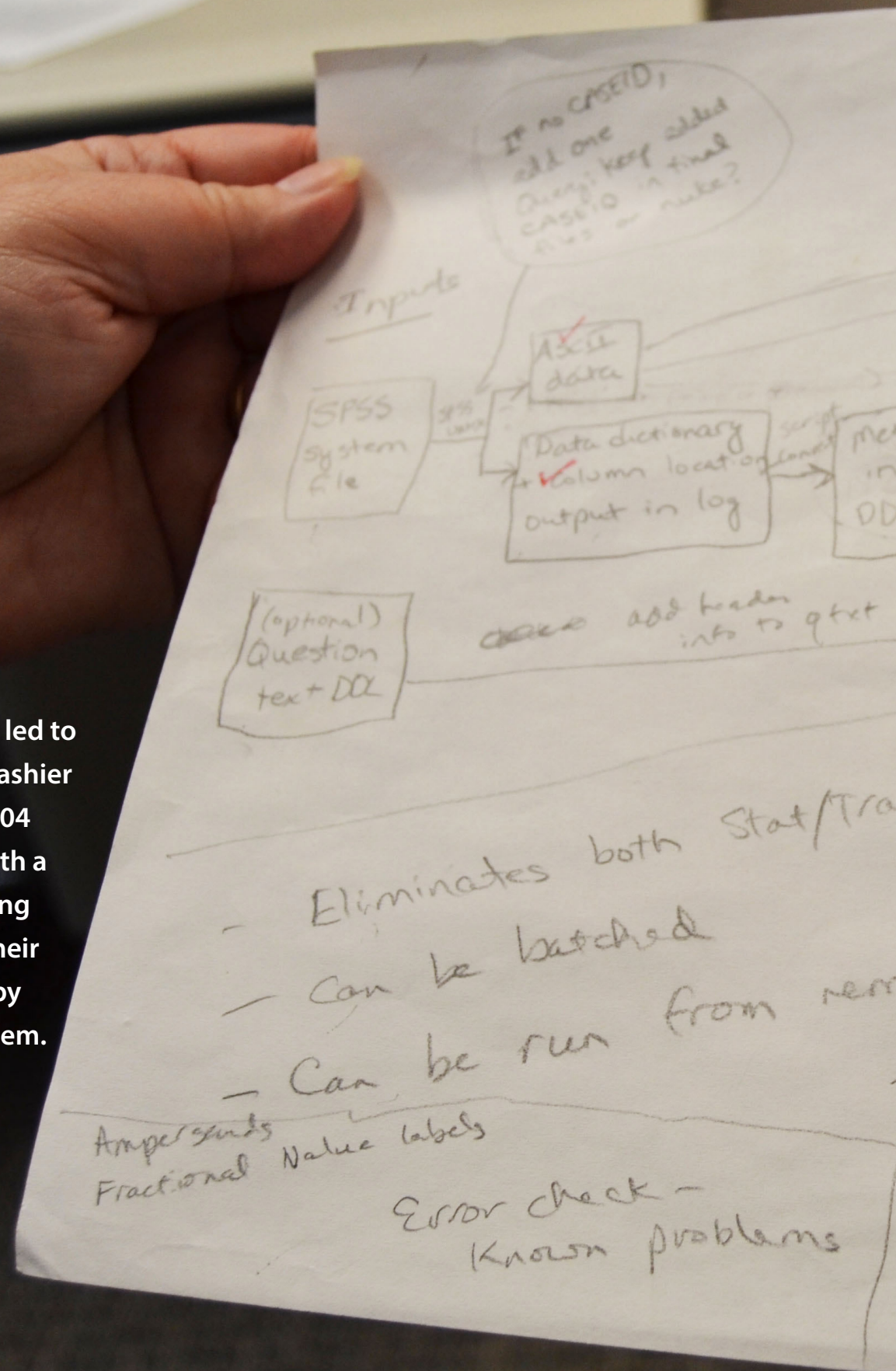


Doctor Ashe

HOW HERMES TOOK FLIGHT AT ICPSR



This is the sketch that led to Hermes. Peggy Overcashier made a diagram in 2004 after talking in jest with a colleague about ridding ICPSR processors of their technical headaches by writing their own system.



Peggy Overcashier

We salute Hermes creator Peggy Overcashier, an ICPSR innovator. Recently, Peggy and some of her peers took time to reflect on how a tool called Hermes helped ICPSR improve data processing and offer data users more options. Here's a recap of those conversations.

ICPSR is renowned for its long history of data stewardship. Much of the credit for ICPSR's success is owed to the group that has helped make more than 8,900 studies and 65,000 datasets available to ICPSR members and the public: the ICPSR data processing staff. These individuals diligently work to improve data quality, make detailed codebooks, and translate studies into multiple statistical packages, maximizing access and ensuring long-term preservation.

An important milestone in ICPSR's data-processing timeline is the creation of a tool called Hermes, named after the winged messenger from Greek mythology. Hermes has allowed data processors to, with a single command, manage an unlimited number of data collections and convert datasets into multiple statistical packages (SPSS, SAS, Stata, etc.).

Hermes | Continued on page 16



“No matter what we did before Hermes, we had to do a lot of it by hand. When you have 20 people processing by hand, they will have 20 ways of putting the data out. So, one question was how do we standardize our products? We’re ICPSR. We needed to have a typical standard so people know it’s ICPSR’s version of data. We had a desire to do more, faster, accurately. Once Peggy went from being a processor to being a CNS programmer, we started to automate some of our processes because she had the experience of knowing what needed to be done.”

Mike Shove, Archive Manager,
General Archive

Life before Hermes

If anyone thinks data processing can be tedious now, they have no idea how it was back then. Before Hermes, ICPSR processors had to do everything pretty much manually. Part of processing is having a master file, which is used to create other products that we distribute (SAS, SPSS, Stata, etc.). Back then, we had to take every file and put it through some mind-numbing, repetitive processes. Before, a change had to be made in every single version of a file or offshoot. Now, if you make a change, you make it in the master, then tell Hermes to change it across packages. It can go through potentially hundreds of files — it’s unlimited, basically — through one single Hermes command.

From sketch to reality

In 2004, former Computing and Networking Services (CNS) Director Bryan Beecher and I were working together on the ICPSR Process Improvement Committee. At the time, I was a data processing supervisor. The committee worked on issues with ICPSR’s existing processing software, and we had a list of problems. One day, I must have been complaining about something because Bryan said to me, “So why don’t you go write your own system?” We had a good laugh about that.

But little did we know, Bryan’s question would lead to big changes. Overnight, the ideas had started to gel for me. “Oh my gosh! We really do have all the pieces to write our own system,” I realized. We just needed somebody to put all the goo together and orchestrate it.

I made a diagram of what a new processing tool could look like and showed it to my supervisor, who loved the idea because she saw that it was going to have long-term time and cost savings, and improvements in quality.

With the support of my boss, it was time to work out the details. There were many barriers. I wasn’t on Computing and Network Services staff. I was still developing my programming skills. And, I had many other obligations on my plate. I knew I couldn’t do this alone. I collaborated with a software vendor (UC Berkeley Survey Documentation and Analysis (SDA)) to get past some

software limitations. Our relationship with them helped make this possible.

Hermes is born

In its infancy, Hermes used commercial software already owned by ICPSR to automate batch production of an enhanced suite of ICPSR products and gave processors an alternative to buggy conversion software. Hermes was able to dramatically reduce data processing time. One example in initial testing showed that it took 14 minutes to complete a task that used to take a processor 40 days to complete. With these time savings, processors were encouraged to focus on value-adding tasks and processing more in less time.

There was a lot of excitement during Hermes’ early days, though it did take a while for people to get to the point where they trusted it. During testing, there was a small group of staffers across archives who did troubleshooting over a six-month period. The testers said it made things better, even before it was put into wider production around ICPSR.

Looking back 11 years later, I’m a data quality specialist in CNS. In 2005, I was hired into a CNS position to automate internal business pipeline processes.

I can say that Hermes has gotten better with time. I think it has definitely done everything I thought it would, but it’s been upgraded to do more than we originally anticipated. Over the years, processors have come back to help tweak it to make it better.

It’s been a journey.

“Times before Hermes weren’t awful, but they were limited. When you hear about ICPSR’s curation process — when we clean and add value to the data — Hermes is one of the tools in that process. Today, we use Hermes as a tool to produce other packages, but it’s a good diagnostic tool as well. When you need a change across packages, it’s easy to just rerun it.”

Tom Zelenock, Archive Manager, NACJD



Peggy’s tips for moving forward with tech-related ideas:

- Don’t be afraid to make suggestions, no matter how outside the box they may seem.
- Carefully consider and be able to clearly articulate the answer to the question: “What problem are we trying to solve?” If you don’t know that, none of the rest of this matters.
- Run the idea by others who may have a shared interest in seeing the idea come to fruition and may be able to enhance the idea further.
- Try to develop some preliminary specifications of what it should do; perhaps draw some pictures of what it should look like.
- Don’t focus on implementation details such as using a specific system or piece of software.
- If you still want to move the idea forward, submit it for consideration.

NEWS AND NOTES

Data Carpentry workshop set for Monday and Tuesday, September 28-29

The popular two-day Data Carpentry workshop is designed to teach basic concepts, skills and tools for working more effectively with data in an open and reproducible way. The workshop, which precedes the ICPSR Biennial Meeting set for September 30-October 2, is designed for learners with little to no prior knowledge of programming, shell scripting, or command line tools. Topics include:

- Getting data out of spreadsheets and into more powerful tools — using R or Python.
- Using databases, including managing and querying data in SQL.
- Workflows and automating repetitive tasks, in particular using the command line shell and shell scripts.



RCMD Director John Garcia accepts a distinguished career award in April. (Photo courtesy of John Garcia)

RCMD’s Garcia wins Distinguished Career Award

The Latina/a Caucus of the Midwest Political Science Association presented its Distinguished Career Award to John Garcia at the MPSA Annual Meeting on April 16. Congratulations John!

Paper examines nondisclosure agreements for commercial and administrative data

ICPSR recently released a white paper titled [“Contractual Limitations on Data Sharing,”](#) authored by Alex Kanous and Elaine Brock. The report was commissioned as part of a project on “Building Community Engagement for Open Access to Data,” sponsored by the Alfred P. Sloan Foundation. “We keep hearing about the ‘data deluge’ and all the data that’s out there, but if you look at the kinds of data that the social science community wants to use, most of it is restricted,” said ICPSR Director George Alter. “Economists, in particular, are getting data from private firms under nondisclosure agreements that limit their ability to share data with other researchers.” The report examines the use of agreements limiting the uses of commercial and administrative data in an environment where funders and journals are requiring more open access to data. The authors’ recommendations include more appropriately titling these agreements (Data Use Agreement, Data Sharing Agreement or Data Transfer Agreement) and being very specific about what data the agreement covers. The report also recommends that researchers and their institutions be allowed to retain ownership of any derivative works that come from an investigator’s use of the original data.

ICPSR prepares to host Biennial Meeting in fall

The [2015 ICPSR Biennial Meeting](#) will take place September 30–October 2, 2015. This year’s theme is “Working Together for Effective Data Stewardship.” With more than 17 workshops and sessions in development, this biennial meeting will highlight ICPSR’s core functions, featuring current data collections and data-related tools and exciting new data projects. The meeting will also feature practical approaches and strategies to work with data producers to share data, write data management plans, steward discipline-focused repositories, and locate tools and resources to assist in curating and managing research data. It will also include formal and informal opportunities to network with fellow representatives (attendees) from member institutions and ICPSR staff. The meeting is open to individuals from member institutions and invited guests and speakers. There is no registration fee.



Photo illustration by Jenna Tyson

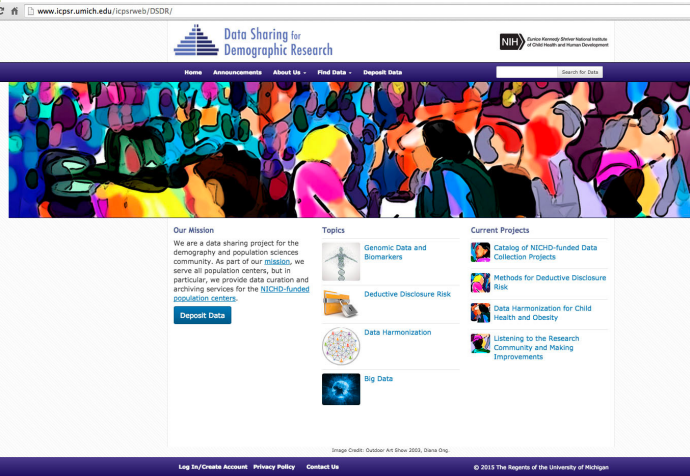
The 2015 ICPSR Biennial Meeting will feature practical approaches and strategies to work with data producers to share data, write data management plans, steward discipline-focused repositories, and locate tools and resources to assist in curating and managing research data.

NADAC website off to a great start

The website for the [National Archive of Data on Arts & Culture \(NADAC\)](#) made its official debut in January. NADAC makes data about the value and impact of the arts available to researchers, policy-makers, arts practitioners and others, with free access provided by The National Endowment for the Arts . To date, NADAC’s most-downloaded data come from Study 35478: General Social Survey, 2012 Merged Data, Including a Cultural Module [United States], with 812 downloads from January 1 through June 1. The second-most-popular data come from Study 35168: Survey of Public Participation in the Arts (SPPA), 2012 [United States], with nearly 200 downloads for the same period.



ICPSR



New website for Data Sharing for Demographic Research (DSDR)

[Data Sharing for Demographic Research \(DSDR\)](#) has released its new website publicly. There are several new features on the website that reflect new directions for DSDR. The new site includes a section that indicates the focus of the new directions for DSDR: the project is enhancing and building is advisory capabilities in genomic data, deductive disclosure risk, data harmonization and big data. Of particular note is the collaborative effort with the School of Public Health at the University of Michigan. DSDR will be actively seeking feedback from the

community on ways that ICPSR could improve its curation and dissemination practices and thus there is a prominent place for feedback along with an explanation of the curation and dissemination process within DSDR and the types of innovative actions being taken to enhance the process.



January 1 – June 1, 2015

Rank	Study	Archive	Downloads
1	National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994–2008 [Public Use]	DSDR	2,635
2	General Social Survey, 1972–2012 [Cumulative File]	ICPSR	1,998
3	National Survey of Midlife Development in the United States (MIDUS), 1995–1996	NACDA	1,126
4	National Survey of Midlife Development in the United States (MIDUS II), 2004–2006	NACDA	1,032
5	Gender, Mental Illness, and Crime in the United States, 2004	NACJD	884
6	General Social Survey, 2012 Merged Data, Including a Cultural Module [United States]	NADAC	812
7	Children of Immigrants Longitudinal Study (CILS), 1991–2006	DSDR	794
8	Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2013	NAHDAP	784
9	Chinese Household Income Project, 2002	DSDR	707
10	India Human Development Survey (IHDS), 2005	DSDR	684
11	National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004–2009	NACDA	673
12	Police-Public Contact Survey, 2011	NACJD	638
13	Law Enforcement Management and Administrative Statistics (LEMAS), 2007	NACJD	585

PROJECT SEAD MOVES TO ICPSR

The Sustainable Environment Actionable Data (SEAD) Project, directed by Margaret Hedstrom, moved to ICPSR earlier this year. Hedstrom is the Robert M. Warner Collegiate Professor at the University of Michigan School of Information and faculty associate at ICPSR.



What is SEAD?

[SEAD](#) is a National Science Foundation-funded project to create data services for sustainability science research. SEAD provides project spaces in which scientists manage, find, and share data, and connects them to repositories (including ICPSR) that will provide long-term access and preservation of data.

What makes SEAD and ICPSR good partners?

“There is this notion of the data life cycle in SEAD, and a strong focus on curating data, which resonates with the work of ICPSR,” said SEAD Program manager Sandy Payette. ICPSR Director George Alter adds, “ICPSR has been a partner in SEAD from the beginning, and Margaret and I believe that moving SEAD to ICPSR will have many benefits. It’s a wonderful fit because they’re really doing complementary work.”

How does SEAD impact the data life cycle?

SEAD’s technology supports Project Spaces for researcher teams that allow for collaboration and sharing of data and testing of interactions with applications and computational environments; a Virtual Archive; and a Research Network that includes profiles, citations and publication links.

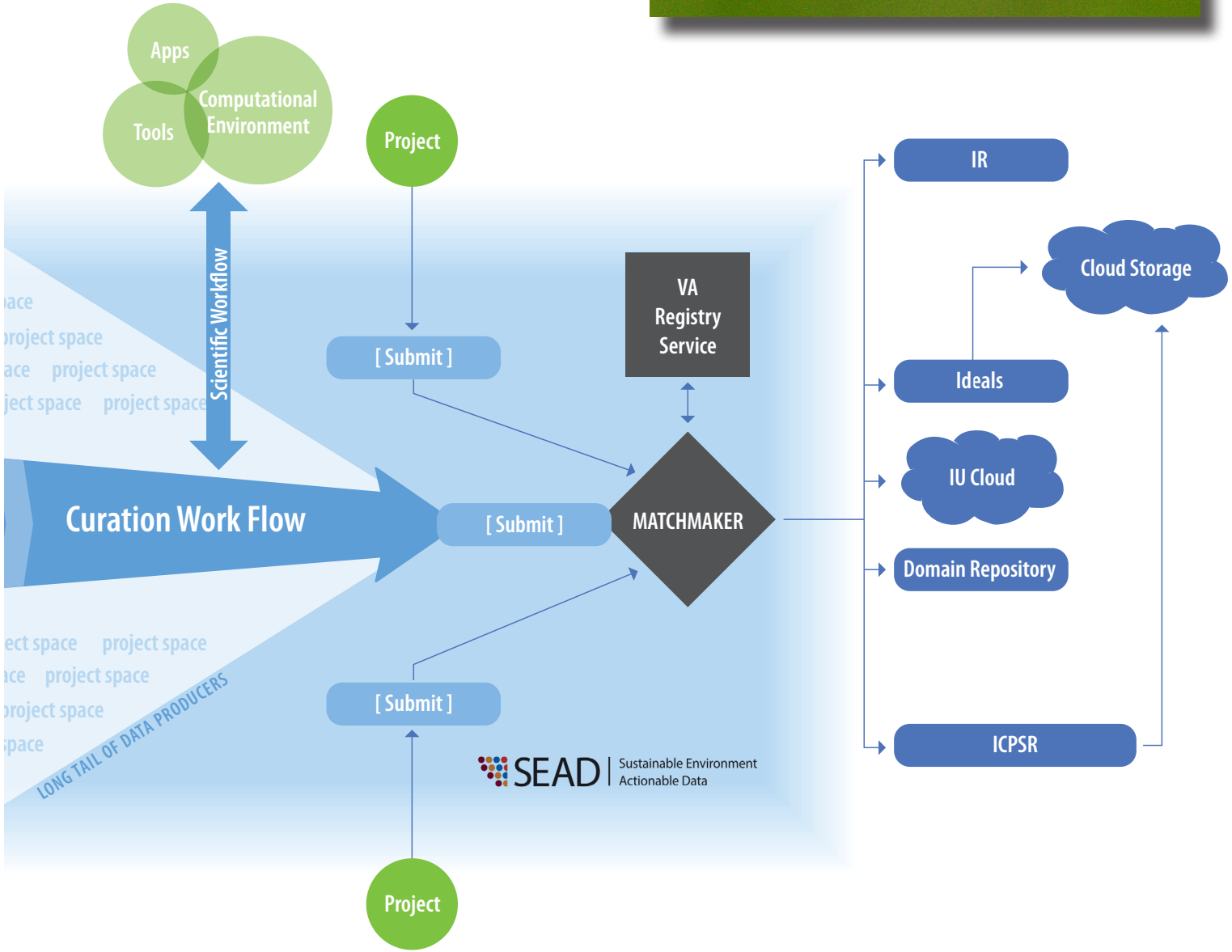
Payette describes a process in which SEAD directs researchers through each step of the data life cycle. “SEAD provides the ability for scientists and project groups to collaborate, submit data, and curate data in an active Project Space. Then, the data can move to the next state of existence, which is a published state, where the data itself can become a publication. Within the SEAD environment, when that happens there’s a component known as ‘The Matchmaker,’ associated with the Virtual Archive. That’s the point where rule-based action can occur, which is, ‘where is the best place for this data to go?’ ICPSR is one of the target areas,” Payette said.

Who does SEAD typically serve?

- Researchers who produce and analyze heterogeneous data that is unique and at a fine resolution and granularity
- Researchers who need to collaborate with others to make scientific advances
- Researchers who lack access to reliable cyberinfrastructure for managing, sharing, analyzing, publishing and archiving data “This is a notion of a national, emerging, data infrastructure,” Payette said.

Next steps?

“As research in the social sciences becomes increasingly complex, SEAD provides a new set of tools to manage collaboration and data sharing,” Alter said.



SOCIAL SCIENCE MEETS CRYPTOGRAPHY

A project led by ICPSR Director George Alter aims to develop a tool to enable joint statistical analysis across multiple private datasets, without the need to share underlying data. Two meetings recently at ICPSR introduced a community of social scientists and statisticians to a group of computer scientists and cryptographers who are experts in secure multi-party computation (MPC).

Imagine that you are a researcher who wants to analyze data from three competing companies, but the companies don't want to reveal data to their competitors. What if there were a way to conduct your analyses without the companies sharing their data with you or each other?

Earlier this year, ICPSR Director George Alter introduced social scientists and statisticians to a group of cryptographers who are experts in secure Multiparty Computation (MPC). MPC is a branch of cryptography that allows joint statistical analyses across multiple private datasets without sharing the underlying data.

One outcome of the two meetings is a roadmap to make the project's vision a reality, which will be described in a forthcoming white paper. "What was needed was for the two communities to meet so the cryptographers could hear the needs of the social science community," Alter said.

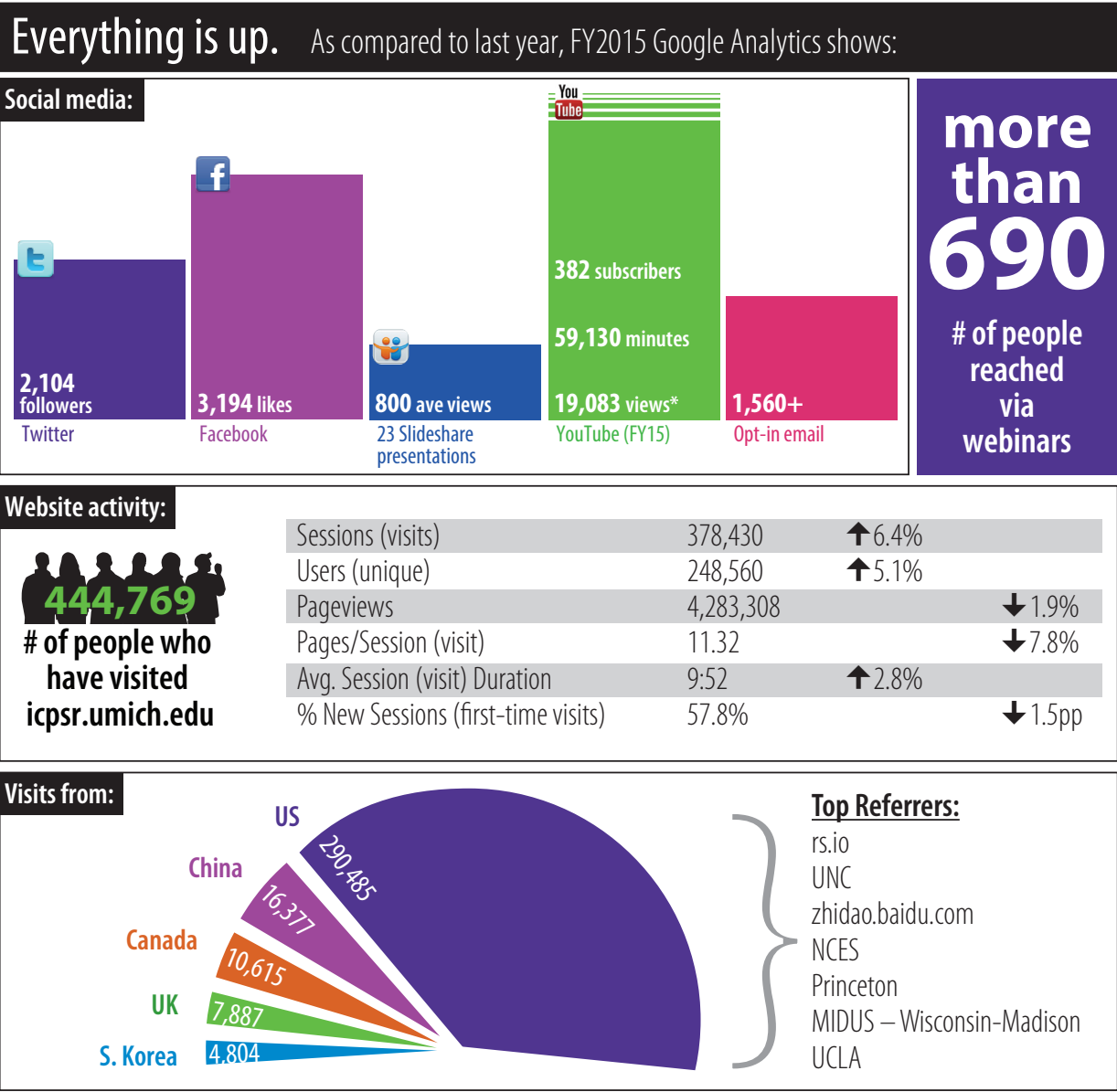
Many government agencies hold data that they don't share with each other, much less share with researchers, Alter said. Sometimes they are prevented from sharing by the statutes under which they operate. It's extremely difficult to get any information that could be associated with individuals, but researchers don't need to know who the individuals are, Alter said. "What we're looking for are patterns, trends, relationships. We don't need to know who the people are, but how they fit together."

In the MPC model, you can combine data from two different databases without ever revealing any individual's information. Database owners know what information they're sharing. They do the encryption, and they hold the key.

So, in theory, there is a way that the companies in the scenario above could still provide researchers the information they seek, while being confident that they are not releasing any of the data itself, Alter said. "Now, we just need to build it."



ICPSR Director George Alter invited cryptographers to a workshop called "Computing Statistics from Privata Data." Story and photo by Dory Knight-Ingram



INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH	Council members	Philip N. Jefferson Swarthmore College	Bulletin staff
George Alter, Director	Christopher H. Achen, Chair Princeton University	Carl Lagoze University of Michigan	Editor: Dory Knight-Ingram dkni@umich.edu, (734) 615-7904
Peter Granda, Associate Director	Marilyn Andrews University of Regina	Chandra L. Muller University of Texas at Austin	Graphic Designer: Jenna Tyson
Mary Vardigan, Assistant Director	Tony N. Brown Vanderbilt University	Ronald Nakao Stanford University Libraries	ICPSR is a center within the Institute for Social Research at the University of Michigan.
Diane Winter, Assistant Director for Administration	Robert S. Chen, Columbia University & CIESIN	William Vega University of Southern California	
	Colin Elman Syracuse University	Linda J. Waite University of Chicago	
	John Fox McMaster University		