The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
June 3, 2020, 12-2 pm ET

Virtual

Participants

Lisa Arafune (CASC)	Deep Medhi (NSF)
Tom Barton (UChicago)	Elizabeth Niswander (Illinois)
Richard Carlson (DOE/SC)	Michael Nelson (Carnegie)
Dhruva Chakravorty (TAMU)	Donald Petravic (NCSA)
Michael Corn (UCSD)	Steve Petruzza (Utah)
Martin Doczkat (FCC)	Birali Runesha (UChicago)
Sharon Broude Geva (UMich)	Andrew Theissen (NTIA)
Ron Hutchins (UVA)	Sean Wilkinson (ORNL)
Margaret Johnson (NCSA)	KC Wang (Clemson)
Padma Krishnaswamy (FCC)	Alex Withers (NCSA)
Joyce Lee (NCO)	Andrew Younge (Sandia)
David Martin (ANL)	

Proceedings

This meeting was chaired by Richard Carlson (DOE/SC) and Vipin Chaudhary (NSF).

Panelists: *Data Confidentiality (Session 2)*

- **Ronald R. Hutchins, Vice President for Internet Technology, University of Virginia**
- **Hakizumwami Birali Runesha, Assistant Vice President for Research Computing; Director of the Research Computing Center, The University of Chicago**
- **Joanna Lyn Grama, Associate Vice President, Vantage Technology Consulting Group**
- **Alexander Withers, Chief Information Security Officer and Assistant Director of Cyber Security, National Center for Supercomputing Applications**

Recap from Session 1 and Panelist Opening Remarks

Ronald Hutchins: More data using in research is going be protected and/or controlled even open data, to ensure reproducibility. CUI – special protections. Also business information needs protection.

Protecting data and making it accessible for researchers

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

Health data – VA, not all university has ability to host protected health data or other protected data.

Birali Runesha: How to protect yet enable research. At UChicago – hosting and working with sensitive data. Approach: looking holistically at entire picture – external as well as data generated at institution; procuring data. Complex. Future: How to scale and service this? Forums like this are critical to understand how other institutions are handling it/ safeguarding shared data.

Joanna Grama: Nuances: need to protect research and IP, while mindful of compliance "floors" (laws/regulations). Implications for big data and analytics, for example, makes this topic complex. If we don't figure out how to protect data and make it useful. If not meet minimum base of acceptable protection, it will be legislated, leading to unintended consequences (e.g., Sarbanes-Oxley)

Alex Withers: Shared resources. Create walled garden environment: researchers also use other resources. Practical issues: may be working on protected data and may use another computing research and using derived, unprotected data. Adept at protecting data and providing authorized access, but no good way to transition/run in mixed environment. HIPAA, etc. does not facilitate realistic daily operations. Private sector researchers often worry about other users. Protect data while providing pragmatic environment that is of use to your users.

<u>Policy/Law and technology</u>

Joanna: Divide between technology and policy good for flexibility? Researchers translate policies into technology implementation; many interpretations

Law never keep up with technology.

- Unintended impact of laws (e.g., Sarbanes-Oxley)
- Flexible, agile approach. Technology-agnostic laws. That state principles
- If address technology with law, Rulemaking processing is negotiated and placed in standards (as opposed to codes)

Implementing laws – subjective and language disconnect with technical folks.

- Have best technology folks implement and document. Diverse environments for implementing CUI, HIPAA. Document reasoning for interpretation.
- Tension between implementing controls while not impeding mission
- Different frameworks, interpretations: Issue of sharing data between institutions that have interpreted data differently

Have live feedback loop between technologists and policy folks – e.g., ask about intent behind policy

HIPAA – sometimes controls. Issues: Folks assume "HIPAA-aligned" data is protected. However, BAA with cloud provider does not ensure protection for end-to-end system. Desire more controls to address security addressed.

Huge Data workshop(just hosted).

Next issue: Using protected/sensitive data for research.

- Unclear tools, policies to enable this use. Technology, policy and people. From past experiences: people running protection protocols is the issue (who and how run protection protocols). How to harmonize : Share clear description of each university protocols.

Joanna:

- Policy always allows sufficient loopholes. People are the random variable.
- Institute for Higher Education Policy use student longitudinal data to understand student success and access. Able to have federal student union record to follow students from Kindergarten on up. Link to wage and tax data, etc., to make findings on educational success.
- Rules of operation – better to come up with these than have these imposed

Alex:

- Automate user-access, training, technical controls is key to this question
- Need to force use of workflow control – but issue of staff resources to implement and maintain

Birali: 2 key ingredients

- SOP and workflow demonstrating process of going from unit to unit
- Develop vocabulary

Ron:

- Virginia universities' Accord project - Shared protected data compute resource for public universities. Office of Sponsored Programs among universities- different languages used among universities.
- HIPAA vs. CUI controls
    - HIPAA: Walled garden provides certain permissions internally, post-access audit
    - CUI: Protected up front.
    - Will be awhile before compartmentalize in clear guidelines
- Business-side protection – complex, not understand entire need

Discussion

Controls in any regulatory control set that affect data security – focus on these. Many controls focus on contractual terms, audits. Tension between data security and obligations of compliance (bureaucratic overhead). Compliance is necessary. How to incentivize next generation approach to audit and compliance that isn't too voluminous. Security controls are find gaps and fill them – anathema to researching culture.

Next gen approach to audit and compliance- concept. How prove can't be breached?

- Ron: Dialogue to build simple system policymakers, implementer and auditors. Need wall to prevent conflict of interest. Also, adds complications Simpler system is more secure
- Alex - NCSA approach: continuous auditing process. Worked on automation. Also facilitates interactions with external auditors. Establish close relationships between security engineers and system admins.

- Birali: Need common interpretation between auditors and implementers. Establish baseline vocabulary and documentation before doing next-generation approach. Connect these worlds.

Open–ended policy and walk through negotiated rulemakings and reach consolidated set of implementation rules (akin to ethernet standards). Like flexibility to interpret based on environment, but need common language to prove doing correct.

Negotiated rulemaking standpoint:

- Opportunity to explain different languages and have a framing discussion. What would happen?
- Common language helpful, but not easy in context of complex, different scenarios and data. Specific needs of each case. Case-by-case approach would be daunting.
    - Establish thresholds between layers – give room for different interpretations within a range. Perhaps go through group of categories.
- Alex: would be great to chat with different institutions regarding different controls. Done with CIS. Space to have a bit more, but would be very challenging. Where to start?
- Internet- lowest levels were flexible and standardized with interfaces and APIs designed to layer diversity on top. Perhaps start with network rules - simplify, build interface?
- Would be great to have next gen approach to audit
- Multiple researchers obtain data from same resource is where some of these conversations can start happening.

Discussion

So far discussed: How to protect data that we have and make it useful for communities authorized to use it –and can meet requirements for access?

In AI context – The more data have, the better our tools will be in doing the analysis. How anonymize data so useful for data without exposing confidential, private information?

- Can company using AI, police itself?
- How research HIPAA data? De-identify data and treat data as if it's confidential.
- Thus controlling access to data despite anonymize, etc., is still needed.
- Imaging and data being de-identified. Working with algorithms, easy to re-identify. AI questioning our SOP – concept of keeping data private. Not need all information.
    - UChicago trying to ensure cannot re-identify. How do you certify de-identification?
    - Technology driving society; need to deal with it now to avoid big, long-term impacts. May be developing too fast without thinking of impacts
    - Need feedback from users of sensitive data and those creating the compliance rules from federal point of view
    - Feedback loop: Developing rules of operation for data governance and access – any law groups working on this issue
    - Rulemaking: reaching the right folks. Also, note notice and comment period does not provide for interaction

Speakers:

ROI and cost efficiency for academic and lab based computing

- CloudBank (Mike Norman (UCSD), Ed Lasowska (UWA)
- PEARC white paper will be ACSM paper (Sharon Broude Geva and Alan Sill (CASC) can discuss issues and also get federal folks.
- NIH STRIDE project – Valerie Virta

Meetings:

June 4: Brussels Forum 2020, The German Marshall Fund of the United States

June 22-25th ISC High Performance opened virtual conference registration

July 26 – 30, 2020, PEARC20 Meeting, Portland, OR (February 17, 2020 deadline for submissions)

Next Meeting: July 1, 2020 (12 noon EDT); Cloud ROI