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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
September 4, 2019, 12-2 pm ET
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

Lisa Arafune (CASC)	Florence Hudson (NE Big Data Innovation Hub)
Wes Bethel (LBL)	Katie Knight (ORNL)
Laura Bivens (DOE/SC)	Joyce Lee (NCO)*
Ben Brown (DOE/SC)	Miron Livny (UW-Madison)
Brian Bockelman (Morgridge)	David Martin (ANL)
Richard Carlson (DOE/SC)*	Shawn McKee (UMich)
Dhruva Chakravorty (TAM)	Michael Nelson (Georgetown)
Wei-Lun Chao (OSU)	Drew Paine (LBL)
Vipin Chaudhary (NSF)	Don Petravick (NCSA)
Kaushik De (UTA)	Nathan Rogers (TTU)
Sharon Broude Geva (UMich)	Birali Runesha (UChicago)
Susan Gregurick (NIH)	Suhas Somnath (ORNL)

Proceedings

This meeting was chaired by Richard Carlson (DOE/SC) and Vipin Chaudhary (NSF). August 2019 meeting minutes were approved.

Guest Speaker

Susan Gregurick Senior Advisor, Office of Data Science Strategy, NIH, *NIH's Strategic Vision for Data Science: enabling a FAIR-Data Ecosystem*

VISION: modernized, integrated, FAIR biomedical data ecosystem

Desired outcome in 5 years, so researchers can scan data and access/analyze it in new and creative ways: ability to link data platforms: link data in Framingham Heart Study with Alzheimer's health data to understand correlative effects in cardiovascular health with aging and dementia

Promise of NIH strategic plan for data science. Imagine:

- Ability to access data in publications (Slide 4, dark data).

[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.

- Linking Data from electronic health care records with personal data and clinical and basic research data (e.g., link to biophysical/biochemical data from expression of proteins, etc. and to contextual data about self and environment)
- New capabilities of AI and advanced technologies to offer medical research, treatment and prevention. Convergence of on-chip computing with AI with HPC or cloud computing; marrying advanced technologies to analyze data large scale, remote sensing or drug delivery

Strategic Plan for Data Science: Goals and Objectives Map (Slide 8). Integrated and collaborative ecosystem working under FAIR principles (Findable, accessible, reusable, interoperable data and tools (slide 12)). Implementation Progress (October 2018- present)

- Data infrastructure
- Modernized data ecosystem
- Data management, analytics and tools
- Workforce development
- Stewardship and sustainability

Implementation Progress (Oct. 2018-Present)

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

NIH developing data management and sharing policy (Slide 13)

- Solicited community input on benefits and challenges (e.g., policy and implementation must go hand in hand, to ensure sufficient infrastructure for researchers to understand and comply)
- Draft policy to be released Early 2020; effective FY2021. Data that underlies publications must be fair and accessible

Overview of sharing publication and related data (Slide 14)

First choice: open access data sharing repositories is first choice (see link)

Other options:

- PubMed Central – up to 2GB and use global identifiers for supplementary materials
    - See Associated Data: part of supplementary data and global unique identifier (Slide 15)
- Commercial and non-profit repositories
- STRIDES Cloud Partners (Google and Amazon) – very early phase of storing and managing large scale high priority NIH datasets

NIH Pilot with Figshare: for data lacking a home

- To understand scale of landscape of amount of potentially shareable dark data (Slide 16)
- All for FAIR implementation of data

- All data submitted has DOI to support data citation, research and metrics
- Support data platform that is openly available and will be exported to other NIH systems
- Goal (where will we be in 1-2 years):
    - better understand data repository landscape (Slide 17) – location of gaps, relationship between domain-specific and generalist repositories (e.g., Figshare)
    - Understand their usefulness in order to strengthen FAIRness of all data repositories
    - Why: To make it easier for researchers to more easily share, find and reuse data

Connecting NIH Data Ecosystems

STRIDES (Science & Tech Research Infrastructure for Discovery, Experimentation and Sustainability Initiative) (Slide 20):

- Google/Amazon p'ship to provide discounts on cloud storage and compute
- Lower barrier to cloud
- Dataset examples (NCBI data resources (12PB!) -largest biomedical dataset in cloud environment (Slide 21)
- Data Analytics using STRIDES Cloud ((Slide 22): AI data algorithms at scale (12 PB or larger); inference of data anomalies

NIH Data environments are rich, but siloed (Slide 23) (e.g., Kids First, data STAGE, AnVIL)

Connect these resources such that NIH investigator can access? (Single Sign-on Across NIH Data Resources) (Slide 24)

- Model for a distributed world (exploring federated authentication) (Slide 24 – 25)
    - Looking to authenticate based on era identities and also ORCiD, other protocols
- Authorization – depends on resources; build translational resources to weave in authorization protocols per resource and translate across (Slide 26-28) to understand user, role and data set access through token process. Developing Minimum Viable Product addressing authentication, authorization and standardized auditing and logging protocols
- Standards -Based approach to authentication, authorization and auditing/logging (Slide 30); also driven by data access policy

Engaging with Broader Research Community (slide 3)

Adopting and expanding on FHIR (fast, healthcare, interoperability, resources) standard and application program interface;

- FHIR designed to exchange electronic health care record data between health care providers and insurance companies
- NIH Guide Notice issued – encouraging use of FHIR in basic and clinical research
- Notice of special interest to SBIR/STTR applicants – interested in receiving applications by small businesses to develop applications based on FHIR standards
- Why? To take advantage of amount of data generating now and in future
- Key applications in AI (Slide 34)

- NIH funding areas in ML (Slide 35): many applications involved image analysis, systems pharmacology, or precision modeling and early detection/screening of cancer and other tumors

AI: Legal and Ethical Challenges- AI working group of advisory council to director

- no clear rules for consent in data use;
- privacy threat;
- potential for bias and discrimination due to training sets that are not completely inclusive of populations studied
- data misuse
- Charge: understand opportunities for AI in NIH; how NIH can build bridge between cs community and biomedical community
- How facilitate training marrying biomedical research with cs
- How identify and understand some ethical considerations in using AI to develop these algorithms
- Themes:
    - making AI data more readily available to researchers;
    - help researchers bridge gap between biomedical and CS; i.e. improved multi-lingual ability;
    - ELSI (ethical, legal and social implications) (Slide 39-40)
    - important areas to apply AI and to advance AI
- Need training sets to include underrepresented and marginalized populations, etc. (Slide 40)

Workforce Development

- Graduate Data Science Summer Program (13 master's level interns for 2019)
- Pilot driven by discussion with local universities
- Coding it Forward: 9 undergraduates from cs placed in NIH Institute and Centers – worked in administrative offices and centers (- worked on automated programs to onboard researchers into STRIDES, etc.)
- Data Science Senior Fellowship – data science and technology experts (large volumes of biomedical research data, impact public health, gain policy exposure)

Discussion

- Who pays for moving data out of cloud? Negotiated egress fees for universities; higher for PIs. Need to pay for computing or for moving data out.
- Dark data: Want future program accessing data from publications. NIH doing through PubMed Central and NCBI. Having data deposited early in the process makes easier for data quality, appropriate MD in tags. UMadison doing some work with DARPA on automated extraction of information.
- AI: ethical use of data is critical. NIH safeguards to avoid unethical use of data? NIH: working group phase of understanding scope and scale of problem. No policy currently. NIH has safeguards re: data access. Incorrect inferences; training sets. TAM available to help.

- NIH partnering with Office of behavioral and social science research to understand social implications of AI, technologies through more funding, research, workshops and Advisory Council to the Director.
- Sign on, authentication, authorization: offering these capabilities for users.
    - Problem addressed from perspective of having interactive user. Makes easier for workflow systems to automatically access data. Time out issues with automated workflow processes; while authentication for users is not that difficult, authorization is because requires nonstatic system.
    - NIH in early phase of mapping out this process. Waiting for initial pilot before mapping potential solutions; stay connected to be in touch in future
- UN forum addressing AI and internet governance – ethical implications is vague. NIH more narrowly focused on ML. Other hot topics: making data sets AI ready? Meaning? Appropriate metadata, etc.? Working with community; want to avoid unintended inferences that are errors in understanding. Charter for NIH working group to be sent to group (see https://acd.od.nih.gov/working-groups/ai.html)
- FY21 OMB R&D Priorities Memo: leveraging power of data and separate subsection on biomedicine; public/private partnerships re: medical health data

MAGIC Annual Planning Meeting (APM): (Incorporates August minutes, September discussion in RED)

Current FY20 topics: Data life cycle series, single session on a range of topics

FY21 proposal: Individual topics instead of multi-month series; lay out 6-8 topics and assign to someone to identify potential speakers

Data integrity for scientific endeavor within data life cycle context (data provenance, security)(F. Hudson)

- Focus of NSF Cybersecurity Center of Excellence (invite Von Welch or Jim Basney)
    - Data integrity, etc. within data life cycle (working on encrypted data (secure nodes); sites setting up data enclaves associated with HPC centers)
- Wide range of topic areas: from working on encrypted data (secure nodes); sites setting up data enclaves associated with HPC centers

Data confidentiality (National Cybersecurity Center of Excellence (NIST); note NIST proposals requesting input on identifying and protecting against breaches and How to recover from breaches)

- NIST proposals requesting input: Data confidentiality – identifying and protecting against breaches and How to recover from breaches
- Interesting topics? Something new in data confidentiality? New data confidentiality projects

Implications of new AI to science work (workloads for HPC ecosystems (identify needed computers and middleware, network structures, etc.)

- DOE running series on AI for science (need summary)
- AI for optimizing HPC – work on IO systems
- Annotating data where annotation environment is difficult; multi-domain (e.g., earth science and medicine); viewed as infeasible by investigators (Don Petravick)

- Example: 4 or 10 images to consider simultaneously to decide if something is a cloud; parallels to medicine – determining whether cancer and not in 1 image. Need to classify. In practice, it's a bar to science; i.e., can't get what you need for starting supervised learning.
- Data set that needs to be partitioned. If no ground truth, need to go through process to find out what data represents for small fraction of it, to build classification engine and classify remainder
- Ground truth for supervised learning hard to obtain, particularly when go beyond RBG (20-30 input channels). Subtopic of AI is difficult classification problems – getting something for supervised learning. How do you not exhaust classifier?

MLAI, Data science, virtualization and containerization approaches -context of workflow (Dhruva Chakavorty)

- Latest occurrences in these areas. 20 min overview from someone working in all 3 areas or workshop?

ROI and cost efficiency for academic and lab based computing (A. Sill)

- 2 PEARC papers led by Craig Stewart (Indiana University) – calculating ROI quantitatively and financially and in human terms.
- CASC: discussing topic of optimizing cost efficiency in delivering computing is a sensitive topic, but need to address it. Can't have intelligent conversation about use of cloud computing without thoroughly examining this topic. – invite Craig, others to review results

Direct integration of energy sources and computing facilities (A. Sill)

- Multi-disciplinary - integrating energy production and computing more closely. Many startups locating data centers remotely, near sources of renewable energy
- Zero carbon cloud: Andrew Chin working on feasibility of stranded power to power small data centers
- Training is requiring larger and larger machines – using HPC resources do conduct training. Training is power intensive

Re-examine networking infrastructure underlying middleware

- Attempts to put more middleware in network layer (authentication, encryption, security)
- Major network providers could discuss what doing to support high bandwidth distributed computing
- Putting storage into network (Miron Livny)
- NRP- latest developments (Vipin Chaudhary/Kevin Thompson)

Next steps:

Rich Carlson, Vipin Chaudhary and Joyce Lee will put together list of topics for 8 months and go back to group for speakers, starting with those who suggested the topics.

Roundtable

TAM: Dhruva Chakravorty

TAM Institute for Data Sciences bootcamp /data science research focus: 120 attendees, including 27 folks from minority serving institutions

Meetings

September 23 - Trusted CI Webinar, NSF Cybersecurity Center of Excellence; "Jupyter Security at LLNL with Thomas Mendoza" (https://trustedci.org/webinars)

September 25-27- CASC meeting, The Alexandrian, Alexandria, VA - will discuss topics relevant to MAGIC's discussions

Nov 1, 2019 -CSSI deadline

November 14, 2019 OAC Core

November 19, SC19, Colorado Convention Center, Rm 711 (1:30 – 3:30 p.m. MT)

Next Meeting: October 2 (12 noon ET)