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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
July 3, 2019, 12-2 pm
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

Kathy Austin (TTU)	Sridhar Kowdley (DHS/HQ)
Francine Berman (RPI)	Joyce Lee (NCO)*
Laura Biven (DOE/SC0	Margaret Johnson (NCSA)
Ben Brown (DOE/SC))	Yuya Kawakami (Grinnell)
Richard Carlson (DOE/SC)	Mike Nelson (Pobox)
Dhurva Chakavorty (TAM)	Don Petravick (NCSA)
Sharon Broude Geva (UMich)	Steve Petruzza (Utah)
Kaushik De (UTA)	Hakizumwami Birali Runesha (UChicago)
Dan Gunter (LBNL)	Mat Selmici (UW-Madison)
Florence Hudson (NE Big Data Innovation Hub)	Arjun Shankar (ORNL)
Shantenu Jha (Rutgers)	Alan Sill (TTU)
	Sean Wilkinson (ORNL)

Proceedings
This meeting was chaired by Richard Carlson (DOE/SC). June 2019 meeting minutes were approved.

Speaker Series: Data Life Cycle

- Dr. Francine Berman, Hamilton Distinguished Professor of Computer Science, RPI, Co-founder, Research Data Alliance, *Organizational challenges to promoting data sharing, stewardship and preservation*

- Hubertus van Dam, Application Architect, Brookhaven National Laboratory, *Online data analysis of molecular dynamics simulations for exascale computing platform*

Data Life Cycle Series Planning
Francine Berman- *Organizational challenges to promoting data sharing, stewardship and preservation*

What problem are we trying to solve? Why care about data sharing, stewardship and preservation?

[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.

- No matter the question of concern, data is an effective and increasingly critical tool to address problems. Data itself becomes what research depends on.
- Need to be able to find data and know what it means.
- Need ecosystem (entire set of tools and structures) to make that data useful; i.e., in the right form for analysis
- Data needs to be preserved for results to be reproducible.

Challenges must be addressed for effective data stewardship and data sharing
All levels must work together, from user level on up, just like ecosystem needs to work together.

Stewardship and Preservation in industry, academia, government: differences in stakeholder alignment
Stakeholders in organizational infrastructure may/or may not be aligned, depending on sector. The more alignment between stakeholders, the better.

Private industry: Stakeholder alignment is not a problem; other data problems
Typically not a problem for commercial stakeholder as it is for academic institution.
Data helps in competitive advantage.
Typically own rights, preserve data for business value and pay for the infrastructure.

Academia: Very little alignment– stewardship, preservation, and data use is problematic.
Stakeholders in universities generate, benefit, own/have rights, preserve or pay for infrastructure
Researchers may generate data, but scientific and broader community benefits.
Grantees generally obtain rights; institutional owners have rights. Unclear who preserves data.
Funders pay for generation and curation of data during grant; but institution may not pick up costs.

Government: Keeps some data, and pays for data that keeps, but subject to politics/leadership, etc.

Why stable organizational support for stewardship and preservation is a hard sell for academic institution?
Care and feeding of data is critical for research.

3 Specific Challenges

1) Fear of Economic Commitment

- When ran SDSC: Goal to promote data ecosystem. How folks with data intensive application get the most out of machines, tools, storage and access and preservation facilities.
    - Provided free repository for users; collected approx. 100 data sets (Graph, Slide 7); became too expensive due to maintenance and upkeep of ecosystem. Graph:
        - Blue line: growth in users' expectation followed rate of growth of data.
        - Red line: SDSC attempt to have sufficient capacity in data storage for users (both tape and disk provided)
        - Note that this is infrastructure, not research. Data stewardship, preservation and tools - view as infrastructure maintenance; more aligned with use.
- Economic Sustainability: Arabidopsis Information Resource Success Story
    - Online model organism database supported by NSF (1999-2013)
    - Alfred Sloan Foundation helped develop sustainability model of subscription.
    - Sustainability is significant issue: need to worry about future of datasets

- Keep everything forever? What data is of value? Reasons change over time and differ by dataset
- Value of data: amount of stewardship investment and for how long? (Slide 10, diagram)
    - Data collection: Value to community (need preservation criteria) and re-evaluate value over time and to individual (assess regularly).
    - Formal institutionalization is important

2) Newsworthiness Problem: Moonshots vs. Disasters (Slide 11)

- Moonshots: Good for research to have great results/breakthrough, not for data infrastructure
- Disasters: are only newsworthy for data infrastructure.

3) Misalignment of stewardship and preservation support with the usual incentive structure (Slide 12)

- Stewardship and preservation support – incentive structure not fit; no recognition for infrastructure as it does not advance research reputation of funding agency, university, researcher
- Researchers are small market and researchers often need custom solutions

Challenge: Organizational Value proposition and metrics of success (Slide 13)

Challenging to advocate for effective data infrastructure to stakeholders

- Measuring success: difficult to quantify lost opportunity/ cost of lack of research infrastructure;
- How determine effective infrastructure or gauge the impact of infrastructure

Towards a solution to the lack of adequate stewardship and preservation infrastructure (Slide 14)

- Move from research to infrastructure and make it sufficiently important
- Identify level of research vs. infrastructure (identify effective ratio).
- Everything should have sustainability mode that makes business sense; focus on what is valuable.

Hubertus van Dam - *Online data analysis of molecular dynamics simulations for exascale computing platform*

CODAR: Center for Online Data Analysis and Reduction (Slide 2, Table)

What CODAR is trying to do (see table):

- Machines retired/o be retired soon (Edison, Tilan, Mira)
- Currently running (Cori, Summit and Teta)
- Up and coming (Permutter (Frontier, Aurara- first machines to have exaflop capabilities))
- Peak performance of these machines and rate at which can store data on these machines (ability to do IO and access data)
- Over the years, machines' compute capability grows faster than rate to access data (see graph). Issue when run simulation, compute data, store and analyze after simulation completed.
- 2 ways: 1) de/compress data; i.e., use substitute compute cycles to do IO; or 2) online data analysis (analyze as produce data without storing)

CODAR with NWChemEx (Slide 3)

Examined 2 applications:

1) Simulate behavior of molecules and calculate trajectories as evolve over time
2) Developing codes for exascale computers and want to analyze codes' performance; worked on extracting data from these applications as applications are running

NWChem Project- Parallel OS Computational Chemistry

- Began in 1990s with idea to design parallel code from ground up;
- Supports wide range of functionality – here, looking at molecular dynamics component of code
- NWChemEx building on reputation of NWChem - re-write NWChem using more modern practices and targeting bigger machines. Using NWCHem as NWCHemEx under development

Use Case: NWCHem MD (biomolecules)

- Transmembrane calcium channel used by plants in responding to adverse environmental conditions.
- To simulate realistically; need to take account of about 1M atoms. Confirmation change take long time. Need 1B time stamps to see change. Need to store data (32PB for trajectory)

Adaptive Sampling Motivation/Challenges (Slide 6)

- Interested only in portions relevant to ongoing research
- Challenges: difficulties in compressing data (atoms - significant vibration at high rate and confirmation changes happen slower). Need to separate time scales.

Batch Approach (Slide 7)

Manifold Learning - use ML approach; looking at intrinsic approach along time line and separates vibrational noise and look at changes in slower modes of motion to compress data.

Depends on atomic positions and forces on atoms (indicates when confirmation change is completed)

Case Study- Different complex (not transmembrane protein): Protein that binds to a DNA and blocks expression of the gene (Video – ran 5k timestamps: DNA moves towards and binds to protein molecule).

- Can do analysis (see graph). After 2k timesteps, DNA meets protein.
- Smoothed angles (graph): Can view important changes (direction change of gradients/forces on atoms).

Case Study- 5 Samples (Slide 9) Batch analysis of data

Selected 5 important points out of 5K timestamp trajectory; Have factor of 1k in compression without losing important data. requires having all data on hand when perform analysis.

Online Adaptive Approach (Slide 10)

Changed to do analysis on fly – using heat trajectory; pumping out and pulling into another application to analyze it. Not all data available up front; must make certain assumptions on what may still be coming. Must adapt algorithm to do this.

Can pick out critical points and get additional points while achieving a factor of 1k in compression and without losing critical information.

Chimbuko- Performance analysis tool that examine the performance of scientific codes as the data comes off simulation (e.g., anomaly detection)

If running parallel application on thousands of cores, and collect information on performance, produce too much data to be stored. Performance data: interested in anomalies.

Chimbuko architecture (Slide 12, diagram)

Diagram:

- NW Chem and Analysis running with communication channel between them
- Instrumented codes with TAU performance tool that collects data on performance events. Pump out and collect data, which goes through analysis tool
- Visualization component: inspects results

Anomaly Detection of Function Executions – (Slide 13)

Event (e.g., incidents occurring at particular time: Entering/leaving with accompanying time stamp). Can be used to reconstruct program's call stack and when subroutines were entered into the stack.

Call Stack Tree (Slide 14)

Important factor in analyzing performance

also includes where in the program the subroutine was called. After a sequence of operations, get different call stacks, same routine executed on different cores and times (get call stack forest). Anomaly detection: examines differences in call stacks in call stack forest.

Online Architecture

Analysis architecture: Code running; emits events.

Analysis of rates that is sufficient to address high volume data:

- Must move much of analysis close to location of data generation. Can't aggregate performance data streams and analyze; move much of analysis to application and perform analysis there.
- Only objects of interest moved to visualization server for further processing.

Chimbuko- Online visualization of front end (Slide 16)

Execution time without/with instrumentation (Slide 18,Graph)

Discussion

Impact of rise of ML on need for data preservation and access; Couple with rise of IoT (highly distributed, decentralized and heterogeneous- what data will be managed and retained). Stewardship and preservation for what?

- ML and IoT calculations and wish to reproduce (need to know where, what and how kept, related rights – policies/privacy issues? Brave new world.

- RDA community: volunteer to build infrastructure together in a way that promotes effectiveness and usefulness. At tip of iceberg. Think about Training sets for AI and importance of transparency, explanability when think of algorithms and bias.
- So much going on with ML calculation under the covers; hard to figure out. Will require sea change in how think about it – how get meaning, what we retain, who does it, etc. No good answer but recognize requires different approach; can't do it same way as before.

Early days, much focus on data collected by satellite sensors. NASA a leader in handling unique data sets. Lessons from NASA experience?

- Many stakeholders align better with NASA– have instruments, data storage; recognize importance. Thus, NASA closer to commercial entities, which helps community stability.
- NASA very engaged with community. Everyone uses data and in different ways. NASA and partners have been thoughtful about standards, which makes a difference in usefulness of data, tools used to ascertain meaning of data.
- From researcher's point of view: get most longevity out of data due to being part of infrastructure, stakeholder alignment and more applicable standards
- Data collected for one mission is sometimes recycled decades later for different purpose. (E.g., old camera images from satellite surveillance – used for soviet test sites; now useful for monitoring climate change).
    - Data Management Challenges – reuse data sites for other disciplines;
    - Findability problem; knowing about existence of data sets; has data changed, etc.
    - Level of unpredictability experienced in data world not accepted in other areas (e.g., electricity)
    - International data sets pose multi-lingual, multi-national challenges
    - Need narrative to support stability in data management and preservation

Success stories:

- Data burst (Harvard committed to data preservation). Useable infrastructure, data burst sites.
- Internet archives
- See incremental path - more realistic (budget, assessment process) – but, difficult to stay the course

From institutional point of view, trends in how institutions view their responsibilities.

- Dealing with conflicting terms; concept of data management plan helped shape thinking of data as "first class object" in research world.
- Agencies experimenting more with funding technical and social infrastructure around data (RDA started because NSF and NIST took risks).
- Problem: institutions difficult to run university as business. Every dollar of revenue – use it as startup package? 1% universities (Harvard, Stanford, etc. have more discretionary funds for experimenting).
- Other universities are more strapped financially; business model not working. Funds more scarce at local level; more important to get "moonshot".
- National level: more awareness that someone needs to be responsible, but universities can't pick up slack.
- Seeking economies of scale; different approaches to reach.

- If get CIOs, libraries together, VPR understands research enterprise as a business and as a discovery enterprise. CIO -> infrastructure. If on same page, seems to work.
- Can continue this discussion in CASC (discuss policy, technology overlaps) to obtain more ideas and thoughts- and report back to MAGIC.

Infrastructure inadequacy for science proposals

- Certain proposals are not offered because impossible or would be declared infeasible
- Current structure: agency requests something generally and researchers try to package accordingly.
- Researchers pitch proposals within a range of reasonableness; i.e., not too risky.
    - E.g., Not where could learn something with known technique, bigger machine or scaled problem size because not reviewed well because not in methodology and technique. Also, not anything that is "outlandish" because won't get reviewed well.

Summary report will be completed after the series ends.

LSN Strategic Plan

Background

MAGIC reports to the LSN IWG, an umbrella group for MAGIC, JET, Broadband R&D Group, JBDT. LSN creating umbrella strategy with input from its member agencies and teams. How MAGIC will fit into its plan is under discussion; probably be a matter of leveraging MAGIC's current activities.

LSN works to coordinate federal activities as part of the Networking Information Technology R&D Program (NITRD). MAGIC is integral part of LSN and needs to be part of strategic plan in how LSN will work in future. Rich Carlson's excerpt from LSN:

> MAGIC would be responsible for agency researchers to coordinate their activities using cloud-based DevOp methodologies (DevOps is methodology doing fast turnaround and looking at future and trying to get best codes-out in best manner and repetitive operations).

Comments beyond multi-month workshop activities and coordinating academic research communities to discuss problems and issues facing research:

- Florence Hudson – Special advisor to NE BD innovation hub (Columbia U) leading EU - US collaboration under Horizon 2020 next generation internet. Meeting in D.C. on July 10. Big data hubs provided joint letter of collaboration to EU group. Part of extended expert group working with the EU. MAGIC is interested in what is going on in international community; Florence will send information to Joyce Lee.
- Rich Carlson: Collaborate with other NITRD IWGs to bring in outside viewpoints to help engage in the data lifecycle issues (e.g., privacy issues in data life cycle)
- Entire data life cycle: perhaps more focus on public access aspect - operations of research and public good should be separate discussions. Look at bigger picture. Not successful as community in terms of implementation (piecemeal approach). Perhaps workshop, etc.? Talk more offline. Perhaps NAAS do roadmap, if applicable.
- Data's aspect as unfunded mandate; assume someone else will be responsible. CASC needs more conversations (best practices and tools). Could invite Fran Berman back.
- Open Storage Network involvement?

Deliverables

Containerization Report: produced by Dhruva Chakravorty (TAM) summarizing February – April 2018 MAGIC Containerization series. He also provided background information. Comments/feedback welcome. NITRD reports will be standardized into a single format.

DevOps Series Report: Seeking volunteer to summarize 2018 DevOps series. Will get credit for summarizing. Joyce Lee will provide presentations, if not online.

Roundtable

July 10, 2019 - Think NEXUS Workshop 2019, Washington, D.C.
September 25-27- CASC meeting, The Alexandrian, Alexandria, VA.
Contact Lisa Arafune lisa.arafune@casc.org) or Sharon Broude Geva (sgeva@umich.edu) if interested in membership.

Next Meeting: August 7 (12 noon ET)