

Committee on Forecasting Costs for Preserving, Archiving, and Promoting Access to Biomedical Data
Board on Mathematical Sciences and Analytics
National Academies of Sciences, Engineering, and Medicine

May 6, 2019
Keck Center of the National Academy of Sciences, Engineering, and Medicine
500 5th Street NW, Washington, DC 20001
Room 106

Open Agenda

Day 1

May 6, 2019

8:30 am –10:00 am CLOSED SESSION—Committee and NAS Staff Only
--

10:00 am – 1:40 pm OPEN SESSION DISCUSSION—DIGITAL DATA ARCHIVING DISRUPTORS

Open session remote login: <https://nasem.zoom.us/j/471850063>

10:00 Welcome, introductions, and statement of meeting objectives

David Chu, Committee Chair

10:05 Disruptors in Digital Archiving: Presentation from the U.S. National Archives

Leslie Johnston, Director of Digital Preservation, U.S. National Archives

Prompting Questions:

- 1) What models do you use to budget for data preservation?
- 2) How do you factor in unexpected cost or budget allocation fluctuations related to data preservation?
- 3) What disruptors have affected appraisal/reappraisal and redaction decisions, how?
- 4) How have those disruptions affected decisions regarding preservation of existing data?
Planning for future data?
- 5) If you employ a cloud-based strategy, what happens if a cloud vendor's services are no longer available?
- 6) How do you think about format obsolescence?

11:05 Disruptors in the Cloud

Vamshidhar Kommineni, Principal Project Manager, Azure Blob Storage, Microsoft

Prompting questions:

- 1) What changes in technologies, data volumes & types, and data uses might appear in the next 5-10-25 years that would be disruptive to cost models and risk assessment for data preservation, archiving and access?
- 2) How do you forecast total cost of ownership of a cloud-based archive based archive over a 5-year life span? Over 10 years?
- 3) What specific steps does your organization take to prepare for any of these eventualities?

12:05 Lunch—available for purchase in the refectory

1:00 Indicators of data management costs at CERN

Simone Campana, Deputy Project Leader of the Worldwide Computing Grid

Prompting questions:

- 1) How does CERN determine what the lifespan of data saved?
- 2) CERN has long time lines and the data generating rate is reported to be 25 PetaB/year. How does CERN plan for storage costs? What is CERN's idea of a planning tool?
- 3) Zenodo - a general-purpose open-access repository - is run "as a marginal activity" - What does that imply for cost forecasting (e.g., how can CERN assume that it remains marginal)?
- 4) How has the archival infrastructure evolved at CERN? How do they expect it to evolve? How open is CERN about its forecasting assumptions?

1:40 Open session adjourns

2:00 pm – 8:00 pm CLOSED SESSION—Committee and NAS Staff Only
--