

Highlights from



Workshop on
Pioneering
the Future of
Federally-
Funded Data
Repositories

National Coordination Office for Networking and
Information Technology R&D (NITRD)

January 13-15, 2021



Carlos Fernando Gamboa
SDCC Liaison meeting, Brookhaven National Laboratory, Feb 4, 2021

Workshop motivation

Why this workshop? Why now?

- Federally supported research data repositories play a strategically important role in increasing access to the results of federally funded research and advancing agency missions, with wide-ranging benefits to research communities, industry, and the general public. They are building blocks of the data science ecosystem.

Strategic Importance



- The data science ecosystem is changing rapidly:
 - Large scale open science cloud infrastructure;
 - New expectations from publishers, funders, and users;
 - AI/ML
 - Convergence of AI, Big Data, and HPC;
 - Changing expectations around privacy, security, and integrity, ethics and bias.

Changing Landscape



- Interoperability requires coordination across community, political, and social boundaries. International counterparts are engaged in large scale efforts in data governance and data infrastructure.

Global Perspectives



Workshop goals and forecasted outcomes

- Long term vision (5-10 year)
- Scientific opportunities and challenges
- Opportunities for cross-agency coordination
- Networking and community building

After workshop

NITRD report publicly available

Workshop organization

Public vs Invited Sessions

Three days
from 1PM-5PM

- This is a Public + Invited meeting
- All Plenary Sessions are public
 - Will run in Zoom Webinar mode
- Breakout sessions on Day 2 and Day 3 are invitation-only
 - Will run in Zoom Meeting mode
- For Plenary Sessions
 - Observer/public participants:** can only use the Zoom Q&A feature to ask questions
 - Panelists/invited participants: can only use the Zoom Chat feature, and/or raise hand, to ask questions and/or to make comments

My role at the workshop

Agenda

- Day 1: Sharing ideas and perspectives
 - Keynote speakers
 - Participant Lightning Talks
 - Announcements from Federal Agencies
- Day 2: Repositories of the Future
 - Opening remarks
 - Keynote speakers
 - Brainstorming breakouts: Envision a more productive, more impactful future for repositories
 - Q: What is the vision, what are the gaps? E.g. What and Why
 - Reports from breakouts
- Day 3: Learning and working together
 - Opening remarks
 - Keynote speakers
 - Brainstorming breakouts: How can repositories, agencies, and others work together to address the opportunities and challenges?
 - Q: What are reasonable action items? E.g. Who and How
 - Reports from breakouts

Only *Invited Participants* contribute to breakout discussions

Everyone re-joins for the report outs on days 2 and 3. Use Plenary Session zoom link

38 Submitted lightning talk abstracts
10 presented

Organizers

- Chaitan Baru (NSF) Co-chair
- Laura Biven (NIH) Co-chair
- Rajeev Agrawal (DOD)
- Frances Carter-Johnson (NSF)
- Ishwar Chandramouli (NIH)
- Wo Chang (NIST)
- Jody Crisp (ORISE)
- Ji Lee (NCO)
- Angela Lester (ORISE)
- Jordan Thomas (DOE / AAAS)

Repositories+ML

Three Case Studies

- The problem of unbalanced datasets
- The problem of cross study validation
- The problem of learning curves

The Problem of Unbalanced Datasets

- A major challenge of data repurposing is that data collected for one purpose (say for measuring trends) might not be balanced for other purposes (say for measuring diversity)
- As more data analysis techniques rely on machine learning data imbalance is an increasing problem
- Imbalance can be from lack of diversity sampling, or from over sampling particular classes, rare events, etc. Could be bias or property of the domain.
- Resampling and other methods can address some but not all of the challenges of imbalanced datasets

Methods for addressing imbalanced data

- Resampling
 - Oversampling
 - Undersampling
- Class Weighting
- Synthetic Data
- Clustering

What Can Repositories Do?

- Periodically process data from the perspective of ML modelers and flag class imbalances
- Produce reports for “most wanted” datasets that would improve diversity and balance
- Provide tools and services to generate balanced subsets of data from unbalanced data
- Gather and publish best practices for user community on approaches to dealing with class imbalance problems
- Open question – Methods for generating **synthetic** data that captures as much saliency as possible

Repositories and Metrics

Kim Pruitt (NIH/NCBI): [NIH Metrics and Lifecycle Work Group](#)

NIH Metrics & Lifecycle Work Group

NIH Data Science Strategic Plan

Modernize the Data Repository Ecosystem

Dynamically measure data use, utility, and modification.

Employ explicit evaluation, lifecycle, sustainability and unsetting expectations for data resources.

Day 1: Sharing ideas and perspectives

Repositories in view of funding agencies

Jerry Sheehan (NIH/NLM): [Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research](#)

All Data Repositories

- | | |
|---|--|
| 1. Assigns unique persistent identifiers | 7. Provides clear use guidance |
| 2. Plans for long-term sustainability | 8. Uses documented security and integrity measures |
| 3. Accompanies data with metadata | 9. Ensures confidentiality |
| 4. Provides mechanisms for curation and quality assurance | 10. Uses common format |
| 5. Offers free & easy access to data | 11. Tracks provenance |
| 6. Contains broad and measured terms of reuse | 12. Provides retention policy |

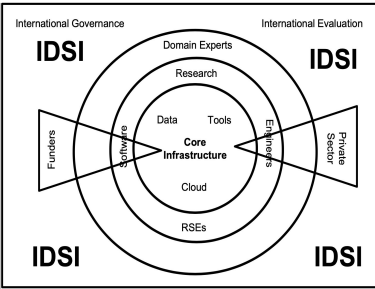
A possible model for repositories

Phil Bourne (University of Virginia): [A SWOT Analysis of Federally Supported Repositories – What Does it Tell Us?](#)

Blue Sky Recommendation – IDSI



- Establish an International Data Science Institute (IDSI) == CERN, Int space station
 - Shared governance, software, hardware, people
 - Cloud based
 - Ten year commitment
 - Measure – Impact, FAIRness, cost/benefit, interdisciplinarity
 - Include all sectors .gov, .com, .org, .edu ...
 - Include all countries



Research Data Framework

Bob Hanisch (NIST): [Research Data Management Framework \(RDaF\)](#)

NIST

MATERIAL MEASUREMENT LABORATORY

RDaF Summary

- Successful in building community interest and engagement
 - Diverse stakeholders
 - National and international
- Challenges
 - Resources
 - Timeliness: the research data ecosystem is changing rapidly. How to keep pace and assure ongoing updates?
 - Controlling scope and scale
- Strategy for moving forward
 - Start with pilot projects in order to validate approach and re-tune as necessary
 - Collaborate with other federal agencies, professional societies, scholarly publishing community, etc., to garner the necessary resources and take advantage of work in progress



Uses AI to find datasets in publications
Found NOAA data sets in 1500 Open Access Publications
Develop usage of Scores

Some Concrete Example of Usage Scores

Diversity: simply count how often the data are used, with special consideration of datasets that are the sole source of information in nearly every study within a research discipline
Landscape-level ecosystems: data that are often used in combination with federal datasets *from other agencies*

Generate scorecards at scale:
Host Competition (Kaggle)

Kaggle's data modeling and analysis platform is designed just for competitions.



5MM+ members



4MM+ uploaded solutions



300+ competitions



50,000+ open datasets

"No matter who you are, most of the smartest people work for someone else."
—Bill Joy Co-founder of Sun Microsystems

RV Guha (Google): Data Commons

Data Commons is an open knowledge repository that combines data from public datasets using mapped common entities.

What we have built

- Open source infrastructure for creating, storing and serving these KGs
 - Storage in conventional relational stores
 - Ingestion pipelines, etc. via Github
 - APIs in Python, REST, R

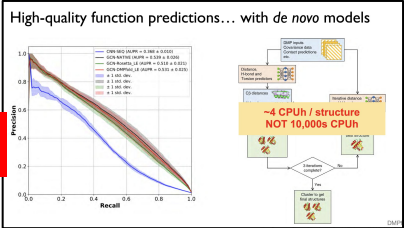
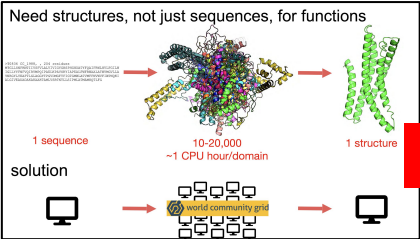
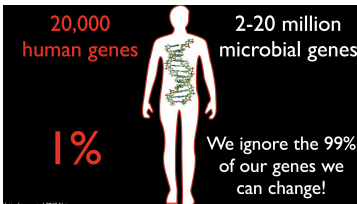
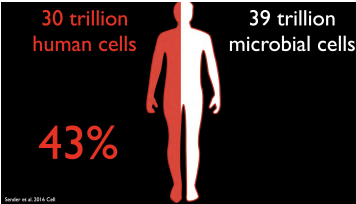
2 large KGs on this infrastructure

- Society KG: demographics, education, crime, ...
- Biomedical KG

Work starting on climate/energy Data Commons

Day 2: Repositories of the Future

Rob Knight (University of California San Diego): [AI Needs to Handle Expanding Microbiome and Metabolome Data Resources—Especially for Global Pandemics](#)



Key needs:

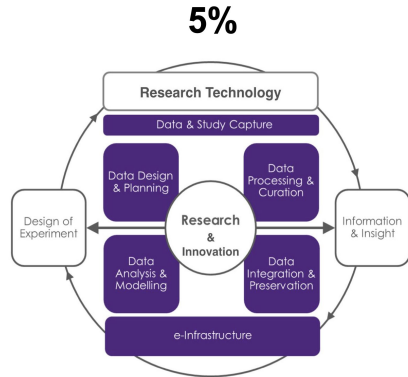
- Data currently in project-specific repositories (e.g. HMP DACC, ISS-MO, NextStrain), generic resources with metadata stripped (Genbank), or under-resourced institutional resources (Qiita, GNPS), limiting reuse
- Metadata not standardized, needs to be hand-curated each time despite GSC and other efforts
- Re-using a particular dataset requires deep knowledge AI and ML algorithms have already had large impact (RF classification and regression, word2vec), lots more potential with good data connections and improved hardware accessibility

Keywords from breakout sessions

- Automate curation workflows
- Repositories for AI
- FAIR and Data quality
- AI and Reproducibility

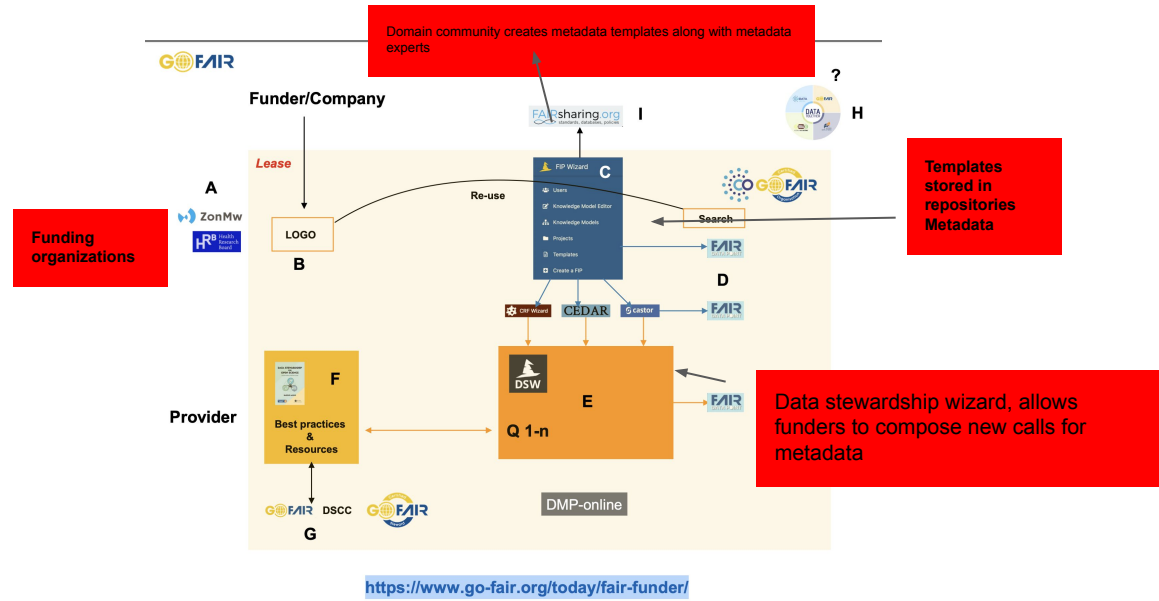
Barend Mons (University of Leiden): "Sorry but no one wanted to pay for the rocket launcher"

Invest 5% of research funds in ensuring data are reusable



GOFAIR: Studied how to allow Funders to realistically mandate, and grantees to realistically produce FAIR research outputs

FAIR research funding cycle



The seven capital sins of Open Science



1: Age factor....Reward only narrative.com

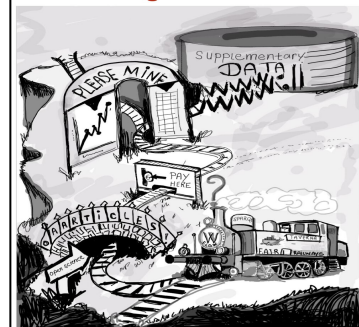


2: Ignore complexity and existing data



Barend Mons (University of Leiden):
["Sorry but no one wanted to pay for the rocket launcher"](#)

5: create a nightmare for machines



7: Create Data without a Data Stewardship plan



3: Disrespect other disciplines



4: publish data without a supplementary article



6: refuse to invest in research -infrastructure



The need to motivate technical evolution of repositories by funding core infrastructure