Opening Science & Scholarship

Michael F. Huerta, Ph.D.
Coordinator of Data Science & Open Science Initiatives
Associate Director for Program Development
National Library of Medicine, NIH

National Academies – September 18, 2017
The National Library of Medicine
The National Library of Medicine

- NIH Institute – Lead, conduct & support research in:
  - Information science
  - Informatics
  - Data science
The National Library of Medicine

- NIH Institute – Lead, conduct & support research in:
  - Information science
  - Informatics
  - Data science

- Biomedical Library – The world’s largest
  - Embrace openness – science and scholarship
  - Steward of literature and more
  - Index > 5600 journals in MEDLINE
  - Major data & info resources
    - Sends > 100 terabytes of data to > 4 million users
    - Receives > 10 terabytes of data from > 3,000 users
What NLM Does

Biomedical Science
Biomedical Science
Biomedical Science
Collections of Digital Research Objects

Biomedical Science

Data

Literature

Software

Models

Workflows
Collections of Digital Research Objects

Curate

Biomedical Science

Information Science
Collections of Digital Research Objects

Biomedical Science

Information Science

Selection
Acquisition
Metadata
Preservation
Classification
Collections of Digital Research Objects

Biomedical Science

Informatics
- Algorithms
- Ontologies
- Computation
- Software tools

Information Science
Collections of Digital Research Objects

Data Science

Extract insight from data

Biomedical Science

Information Science

Informatics
Collections of Digital Research Objects

Data Science
- Visualization
- Statistics
- Probabilistics
- Artificial intelligence

Biomedical Science

Informatics

Information Science
FA

Findable

Accessible

Digital Research Objects
FAI

Findable

Accessible

Interoperable

Digital Research Objects
FAIR

Findable

Accessible

Re-usable

Interoperable

Digital Research Objects
Digital Research Objects

- Findable
- Accessible
- Interoperable
- Re-usable
- Sustainable
- Attributable
Unified Medical Language System
Data Science at NIH
Data Science at NIH

Advisory Committee to the NIH Director Recommended:
Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.
Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.

- **Open Science** - *NLM should lead efforts to support and catalyze open science*, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods.
Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.

- **Open Science** - *NLM should lead efforts to support and catalyze open science*, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods.

*Recommendations Accepted*
Data

Science
Sharing Data  Open Science
Sharing Data  Open Science

Data-centric & open paradigm
Sharing DROs* Open Science

*Digital Research Objects
- Data
- Software
- Publications
- Workflows, etc.
Sharing DROs → Open Science → FAIR
Data-centric & open paradigms have proven successful
NIH Support for Data-Centric & Open Science
Large Scale Data

Human Genome Project
The Human Connectome Project

All of Us
The Future of Health Begins With You

ECHO
Environmental Influences on Child Health Outcomes

FlyBase
MG1

ZFIN
Saccharomyces Genome Database

WormBase

Adolescent Brain Cognitive Development

TOPMed

Cancer Moonshot

NIH Human Microbiome Project

NIH
GTE
The Common Fund

HMP
National Institutes of Health
Data Centric & Open Science

- Requires soft & hard **infrastructure:**
  - Clear & heeded **policies** of funders, publishers, etc.
  - Widely-used data-related **standards** (incl metadata)
  - Data **repositories**, platforms & tools
  - Appropriate **incentive** structure
Data Centric & Open Science

- Requires soft & hard infrastructure:
  - Clear & heeded policies of funders, publishers, etc.
  - Widely-used data-related standards (incl metadata)
  - Data repositories, platforms & tools
  - Appropriate incentive structure

- When implemented with FAIR principles
  - Forms basis of digital ecosystem – transformational
    - Accelerating pace of discovery
    - Changing the nature of discovery
Most domains of biomedical research are neither data-centric nor open.
For these domains, the major public products of research are *scientific papers* that describe the authors’ *conclusions about* the data...
...but the underlying **data are never seen**.
...but the underlying data are never seen.

Much less shared
This is about to change
Societal expectations
Societal expectations

Policy directives
Societal expectations

Policy directives

Technical capabilities
Societal expectations

Technical capabilities

Policy directives

Scientific opportunities
Societal expectations

Technical capabilities

More Sharing & More Open Across All Domains

Scientific opportunities

Policy directives
DataScience@NIH
Build on NIH-Wide Opportunities
DataScience@NIH

Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.
DataScience@NIH
Build on NIH-Wide Opportunities

- **Findable** – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit (10/17)

- **Accessible** - Holdren Memo to increase access
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal
DataScience@NIH

Build on NIH-Wide Opportunities

- **Findable – PubMed**
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.

- **Accessible - Holdren Memo to increase access**
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal

- **Interoperable - Standards**
  - NLM – UMLS, SNOMED-CT, LOINC, RxNorm, etc.
  - Repository & Initiative-related standards across NIH
  - NIH Clinical Common Data Element Task Force
DataScience@NIH
Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.

- Accessible - Holdren Memo to increase access
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal

- Interoperable - Standards
  - NLM – UMLS, SNOMED-CT, LOINC, RxNorm, etc.
  - Repository & Initiative-related standards across NIH
  - NIH Clinical Common Data Element Task Force

- Re-usable – Linking systems of DROs
  - PubMed – publication & data citations
  - NIH data repositories - data
  - NIH administrative systems (info about grants, DMPs, PIs, etc.)
  - NIH Data Commons Cloud – Shared space for compliant DROs, tools, compute, etc.
Pivot to the Future
Pivot to the Future
Strategic Engagement Across & Beyond NIH
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Sustainability solutions – urgent to address
Sustainability

NIH Investment in Data Repositories

$0 $50,000,000 $100,000,000 $150,000,000 $200,000,000 $250,000,000 $300,000,000 $350,000,000 $400,000,000

Sustainability

NIH Investment in Data Repositories

$0

$50,000,000

$100,000,000

$150,000,000

$200,000,000

$250,000,000

$300,000,000

$350,000,000

$400,000,000

Sustainability

Strategic approach may bend the cost-curve
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Sustainability solutions
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Sustainability solutions
  - Enterprise-wide approaches (balance w IC needs)
    - Solve common problems once
    - Lessons learned & best practices
    - Converge on common:
      - Data-related standards
      - Architectures
      - Acquisitions
      - Operational approaches
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Sustainability solutions
  - Enterprise-wide approaches (balance w IC needs)
    - Solve common problems once
    - Lessons learned & best practices
  - Converge on common:
    - Data-related standards
    - Architectures
    - Acquisitions
    - Operational approaches
  - Evidence-based value assessment for investment in policy changes, infrastructure, data acquisition, preservation, etc.
    - Cost vs benefit analyses
    - Develop and use evidence base & models
Pivot to the Future
Strategic Engagement Across & Beyond NIH
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy

- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens
Pivot to the Future
Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy

- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens

- Continue research & innovation in data science
  - Artificial intelligence, analytics, statistics, probabilistics, etc.
  - At-scale curation (metadata, provenance, etc.)
Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy

- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens

- Continue research & innovation in data science
  - Artificial intelligence, statistics, probabilistics, analytics, etc.
  - At-scale curation (metadata, provenance, etc.)

- Strategic incentive structure for data-centric & open paradigm
  - Will require incentives to change behavior of people & orgs
  - Strategically align incentives across ecosystem to maximize impact
  - Likely best done domain-by-domain