NLM and Visual Analytics: A partnership for discovery and engagement

Patricia Flatley Brennan, RN, PhD
Director
National Library of Medicine
NLM Strategic Plan
Transforming Data into Knowledge

Accelerate discovery and advance health through data-driven research

Reach more people in more ways through enhanced dissemination and engagement

Build a workforce for data-driven research and health
Accelerate discovery and advance health through data-driven research
Fostering a ecosphere of discovery

digital research objects
Reach more people in more ways through enhanced dissemination and engagement
New users, New ways

Biomedical & health information access methods & information dissemination strategies
Foster distinctiveness of NLM as a reliable, trustable source of health information & biomedical data
Build a workforce for data-driven research and health
Expand & enhance research training for biomedical informatics & data science
The Human Enterprise: Reaching People Wherever
NLM & Visual Analytics: A Partnership for Discovery and Engagement

- Accelerate use of NLM resources
- Stimulate discovery in VAHC
- Promote sustainability of VAHC strategies, utilities and tools
Accelerate use of NLM resources
The Literature Challenge

27M articles
1M each year
2 per minute
Today’s user search behavior

“The best place to hide a dead body is page 2 of search results”

Most (>80%) clicks happened in top 20 positions.

Over half of PubMed queries return more than 20 results.
What is PubMed Labs?

PubMed Labs is a test site where we are experimenting with new features and tools that eventually may be incorporated in PubMed, in their current or a revised form based on the input we receive. Please try the site and let us know what you think.
Can we do any better than this???
Daily Data Flow at NCBI

The Hubble Telescope generates 10TB of data per year.

- Data Submissions: 10-15 Terabytes (TB)/day
- Internal Processing: 550 Terabytes/day
- Interactive Web: 20 TB/day
- Data Downloads: 95 TB/day
- 200 Petabytes/year

1 Petabyte = 1,000 TB = 1,000,000 GB = 1,500,000 CD-ROMs

GenBank was last released on CD-ROM in 1996
Stimulate discovery in VAHC
RFI

Next-Generation Data Science Challenges

(NOT LM 17-006)
1. Make EHRs better at the point of care
   1. Reduce fragmentation in patient health records and simplify visualization for care givers
   2. Personally-tailored decision systems to help patients
   3. Automate integration of personal data from mobile devices into clinical workflows

2. Extract research value of electronic health records
   1. Algorithms based on patient similarity to drive CER
   2. Real time structuring of unstructured text in electronic health records
   3. Mobile health data monitoring systems that protect PHI
   4. Address bias in health records that are used for research purposes
   5. Nuanced phenotypes including severity, degree and certainty
• Improved ontologies, vocabularies and standards
  – Improved ontologies and crosswalks between information capture systems
  – Common data models and standard vocabularies
  – Create crosswalks among Common Data elements/models
  – Expand personal health data beyond health record to population-level health determinants

• Better, more accessible methods that extend to the cloud
  – Open source analytic and simulation methods
  – Leverage AI and semantic analytics to integrate data sources across cloud services
  – **Technology platforms** to support data storage and analysis by scientists
  – Statistical tools that are more programming oriented than descriptive packages but easy to learn and deploy quickly
  – Develop accurate, privacy-producing linking methods
  – Study how technology has affected documentation of data analysis and information methods

• Curation at scale
  – Automated assignment of standardized metadata to existing datasets and digital files
  – Automatic assignment of metadata early in the research process
  – **Automate integration** of personal data from mobile devices into clinical workflows
Intelligent Learning Systems

- Approaches for engaging users with their health data
- Learner-centric modalities with content ‘chunking’ to modularize declarative knowledge and procedural experiences
- Brain science research focused on learning and retention
Promote sustainability of VAHC strategies, utilities and tools
Do for data what we’ve done for the literature.
Why does anyone care about models?

R, R, R

Rigor, Reproducibility, Reuse
Curator and Custodian: The NLM Collection
What does a library of VAHC tools look like?
RRIDs characterize Methods

- identify the model organisms, cells lines, antibodies, and tools (such as software or databases) you have used
- Blogs, software took kits, algorithm implementation
- include Research Resource Identifiers (RRIDs) within the materials and methods section of their papers

**doi’s characterize objects**

- Complete declarations (objects, articles, books)
- ISO standard, computational registry
RRIDs & doi’s in PubMed
Questions for discussion

1. What does a library of visual analytic tools look like?
2. Who would use it?
3. How does it differ from models, protocols, pathways, etc?
4. How is authority imbued?
5. Should models be stored separately from the data they are used on and the reports they are used for?
6. How can the NLM help advance VAHC?
Reaching NLM

@NLM_news
@NLMdirector

patti.brennan@nih.gov