

Department of Information Culture
and Data Stewardship

Revisiting Data Literacy in the Big Data Landscape

Sheila Corral
scorrall@pitt.edu





Revisiting Data Literacy in the Big Data Landscape

- What does it mean to be data literate in a world of massive open online data and continuous participatory research programs?
- Where should librarians concentrate their efforts to create real value for the individuals and communities they serve?
- How can the profession collaborate to make a difference in our fast-moving data-rich society?



LIBRARIANS'
INFORMATION
LITERACY
ANNUAL CONFERENCE



Overview

- Trends and developments
 - Technology, pedagogy, information literacy and library engagement with data
- Conceptions and approaches to data literacy
 - Social science and science, traditional and contemporary
- The next frontier
 - Big data 2.0
- Review and conclusion
 - Current and emergent data roles
 - The way forward?



Technology trends – the data shift

“...**technology** has enabled **data** to become the prevalent material and **currency** of research. **Data**, not information, not publications, is rapidly becoming the accepted deliverable of research”

Graham Pryor, DCC (2012)

Data is the new currency

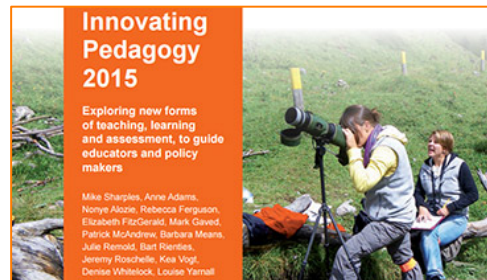
- the new form of exchange in the business world
- + in the public sector where governments are key players



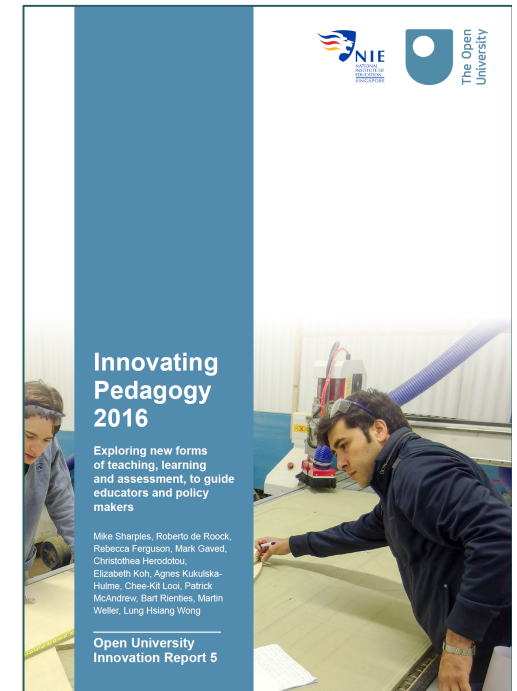


Pedagogical predictions –

- Crossover learning
 - Connecting formal and informal learning
- Incidental learning
 - Harnessing unplanned or unintentional learning
- Context-based learning
 - How context shapes and is shaped by the process of learning
- Learning through social media
 - Using social media to offer long-term learning opportunities
- Learning for the future
 - Preparing students for work and life in an unpredictable future



(Sharples et al., 2015; 2016)



Lifelong and lifewide learning



Information Literacy Development

Reflective discovery

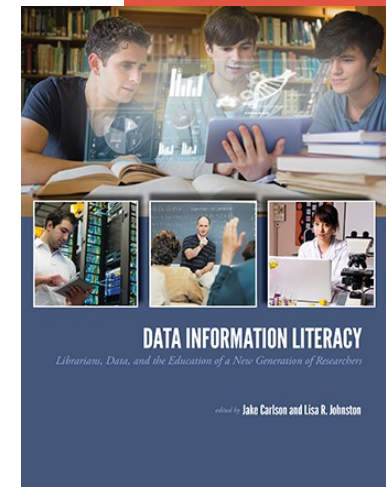
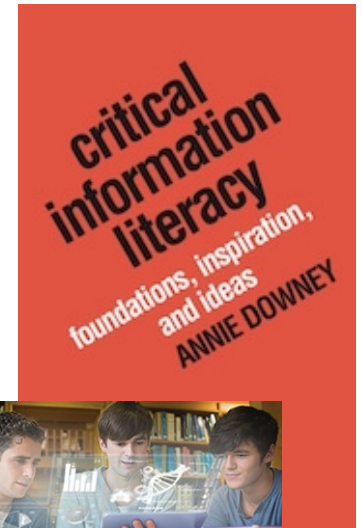
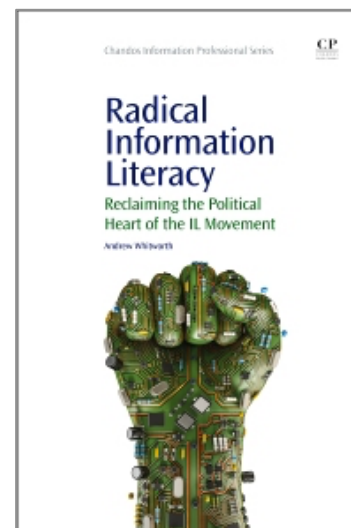
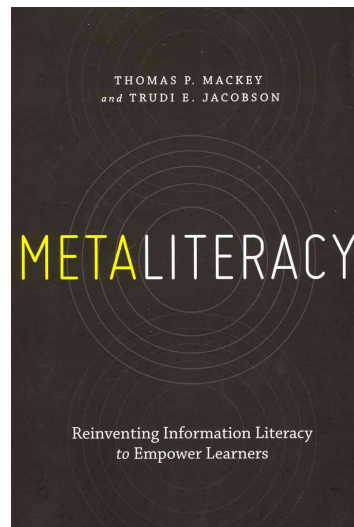
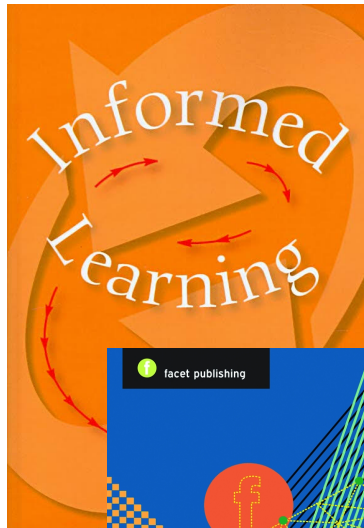
Self-reflexive process

Political, social, and economic issues

Critical pedagogy

Authentic learning

Ethical use





Information Literacy Trends

“Information literacy is for life, not just for a good degree”

Charlie Inskip (2014, June 16)

- Synthesizing the diversity of perspectives and positions on digital, media, information, and other related literacies
 - Acknowledging context is critical for evaluation and use
- Recognizing the multifaceted information needs of learners
 - Using scenario-based assignments and designing literacy support around personal, professional, and academic roles
- Redressing the knowledge, skills, and abilities balance
 - Framing true understanding of the essential concepts

“Navigating information literacy through multiple life perspectives”

(Ruleman et al., 2017, p. 627)



Libraries, Librarians, and Data

The Way We Were (2012) – An Evolving Landscape

- Social science data archives and geospatial data resources
 - local data libraries/support services established in the 1980s and 1990s (e.g., Edinburgh, Oxford, LSE)
- Networked data-intensive science and research data services
 - exploring data curation and storage, advising on data management plans (e.g., Georgia Tech, Purdue, MIT)



(Corrall, 2012)



Libraries, Librarians, and Data

Where Are We Now? – The New Centre of Gravity (2017)

New vocabulary, New roles, responsibilities, relationships

- Data centres/repositories
- Linked data
- Data analytics
- Data warehouses
- Data visualization
- Data journals/papers
- Data citation
- Text and data mining
- Research data services
- Open data for collections
- Learning analytics projects
- Helping researchers to use data visualization tools
- Using data visualization in library assessment
- Metadata consultancy
- Facilitating research using text and data mining

Beyond Research Support

Edward Fox and Virginia Tech researchers earn grant to study big data sharing and reuse



Edward Fox (right) and DAC students.

Congratulations to Edward Fox, professor of computer science and DAC faculty member, who is among a group of Virginia Tech researchers collaborating with Virginia Tech Libraries that has recently been awarded a \$308,175 National Leadership Grant for Libraries from the Institute of Museum and Library Services. The team will be exploring effective ways of storing and reusing big data.

“The IMLS grant will allow contrasting use of the cloud with local infrastructures, like ours that is tailored for integrating focused crawling from the web, tweet collection, collaboration with the Internet Archive, and advanced methods of machine learning, natural language processing, information retrieval, digital libraries, archiving, visualization, and human-computer interaction,” said Fox. To learn more about the grant click [here](#).

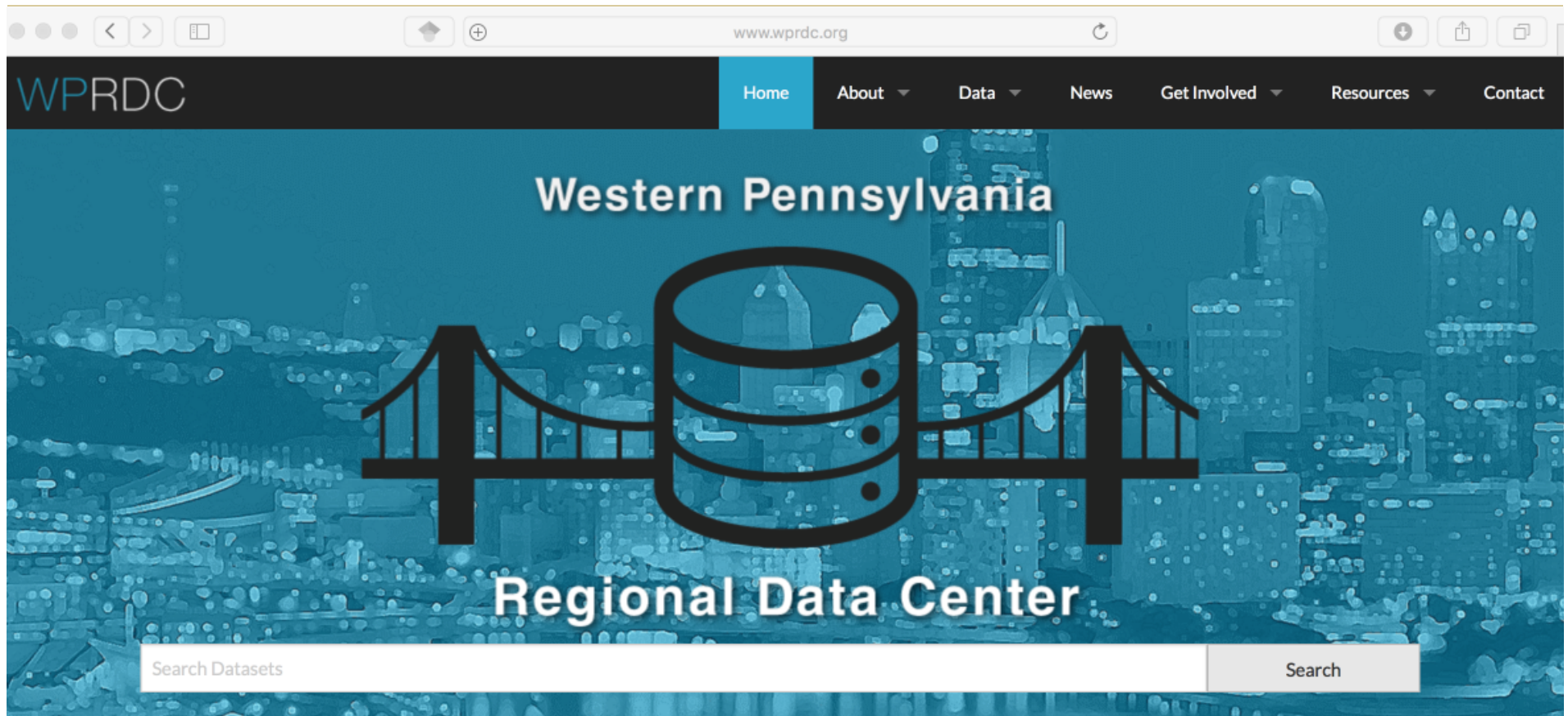


Learning analytics in the library and the emergence of professional ethics conflicts

Big Data has arrived in college campuses and classrooms by way of learning analytics (LA) initiatives. LA measures, collects, analyzes, and reports on student behaviors in order to improve learning environments and outcomes. Now, academic libraries are considering their role in tracking and acting on analyzable flows of student data.

However, LA comes with moral and ethical problems related to students' intellectual privacy and intellectual freedom. ... Libraries must therefore consider how LA initiatives contradict ethical principles set forth in the American Library Association's "Code of Ethics."

(Jones & Salo, 2017, January 19)



Now with New Crime Data!

(Which comes with a handy guide.)

Browse our most popular datasets

156 datasets and growing

Groups/Topics

[Arts & Culture](#)

[Civic Vitality & Governance](#)

[Education](#)

[Business & Economy](#)

[Demographics](#)

[Energy](#)

Organizations

[Allegheny County](#)

[Carnegie Library of Pittsburgh](#)

[Healthy Ride](#)

[BikePGH](#)

[City of Pittsburgh](#)

[Port Authority of Allegheny](#)

Partners

The Western Pennsylvania Regional Data Center is built on partnerships with many organizations. The following organizations and people have been instrumental partners of the Data Center, and we would like to acknowledge their contributions:

- [Allegheny County](#) and the [City of Pittsburgh](#) have been our key partners from the start, and we're excited to be working with so many committed staff and supportive elected officials.
- Our financial supporters include the [University of Pittsburgh](#) and the [Richard King Mellon Foundation](#). The Data Center doesn't happen without their generosity.
- [University of Pittsburgh University Library System Digital Scholarship Services](#) for their help with all things data management and metadata.

About

The Western Pennsylvania Regional Data Center supports key community initiatives by making public information easier to find and use. The Data Center provides a technological and legal infrastructure for data sharing to support a growing ecosystem of data providers and data users. The Data Center maintains Allegheny County and the City of Pittsburgh's open data portal, and provides a number of services to data publishers and users. The Data Center also hosts datasets from these and other public sector agencies, academic institutions, and non-profit organizations. The Data Center is managed by the University of Pittsburgh's Center for Social and Urban Research, and is a partnership of the University, Allegheny County and the City of Pittsburgh.



Conceptions of data literacy [LILAC 2011]

Traditional social science perspective

Data literacy almost synonymous with statistical literacy, quantitative literacy and numeracy – but involving more than basic statistics and mathematical functions

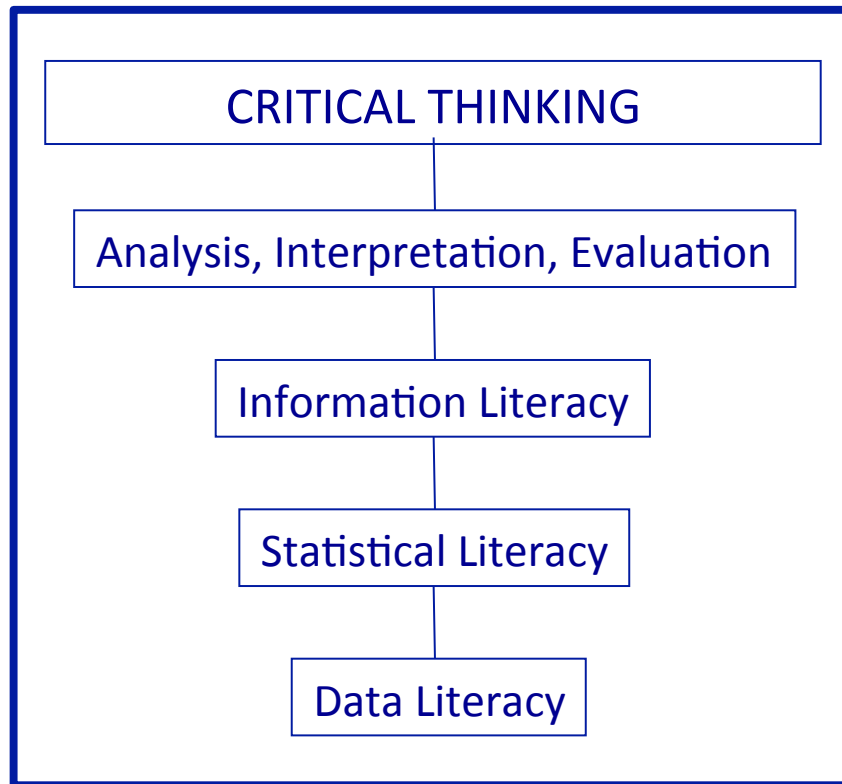
- Understanding data and its tabular and graphical representations, including statistical concepts and terms
- Finding, evaluating and using statistical information effectively and ethically as evidence for social inquiries
- Reading, interpreting and thinking critically about stats

An ***essential*** and ***critical*** component of information competence in social sciences

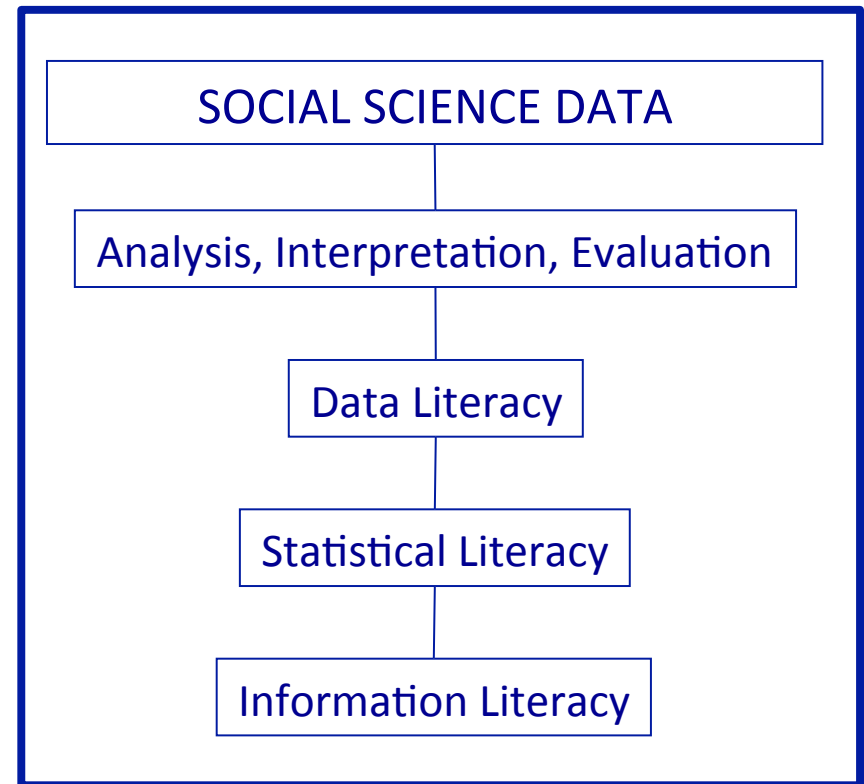
(e.g. Read, 2007; Schield, 1999; Stephenson & Caravello, 2007)



Alternative (hierarchical) social science perspectives



Critical thinking perspective



Discipline perspective

(Schield, 2004)



A STEM/information science perspective

Science data literacy shares aspects of social science conceptions, but requires awareness of the data life cycle, metadata issues, data tools and collaboration mechanisms

- Managing the data generated from experiments, surveys and observations by using sensors and other devices
- Understanding the attributes, quality and history of data to produce valid, reliable answers to scientific inquiries
- Accessing, collecting, processing, manipulating, converting, transforming, evaluating and using data

SDL goes beyond “***pushing***” the data to students by developing abilities and skills in “***pulling***” data

(Qin & D’Ignazio, 2010)



Current approaches to data literacy

- Extending information literacy instruction to research data
 - Initiatives targeting both students and researchers (e.g., Carlson & Bracke, 2015; Doucette & Fyfe, 2013; Haendel et al., 2012; Peters & Vaughn, 2013)
- Preparing graduates for working with data in employment
 - Examples relating to business and public health (Macy & Coates, 2016)
- Reviewing definitions and competencies for data literacy
 - Covering both research data and public sector data
 - Targeting academic, school, and public librarians (e.g., Koltay, 2015; Prado & Marzal, 2013; Schneider, 2013)
- Teaching public librarians about data privacy and training practitioners to empower young people in a data-rich world



Privacy Literacy Training

Teaching NYC librarians how information travels and is shared online, common risks encountered online by users and the importance of digital privacy and literacy.



Led by a team of tech experts, researchers, community activists, and librarians interested in the impact of technological advances on everyone, especially the most vulnerable populations in America, the Data Privacy Project is focused on data privacy literacy, tools, guides, and network building with tech experts to support libraries' increasing role in empowering their communities in a digital world.

This project was made possible in part by the Institute of Museum and Library Services and the Knight Foundation Prototype Fund.

Exploring Data Worlds at the Public Library

[LEARN MORE](#)

About

Exploring Data Worlds at the Public Library

Data literacy is new to the world of libraries and its meaning is still open to negotiation. Although many have advocated for the education of a data-literate population, there is little consensus on what such educational programs should look like, particularly in the context of informal learning at the public library.

The “Exploring Data Worlds at the Public Library” project will investigate youth data literacy in the context of technology-supported programs for young people at the public library. The project aims to increase awareness of the unique data literacy needs of youth as well as to develop strategies for training youth librarians so they can empower young people in our data-driven world.

By addressing gaps in the capacity of libraries to support the public’s knowledge, skills, and practices surrounding data, the project will contribute to IMLS’s priority of building the social and technical infrastructure of libraries nationwide, and to the development of the National Digital Platform—the “combination of software applications, social and technical infrastructure, and staff expertise that provide library content and services to all users in the USA” (IMLS). This project was made possible in part by the Institute of Museum and Library Services. The project’s grant number is RE-31-16-0079-16.



Big Data 2.0 – The next frontier

- Converges e-science with business intelligence, crowdsourcing, big data analytics, social media and Web 2.0 technologies
 - Enables broader and deeper applications of analytical tools
 - Takes data-driven research to new levels of technical and organizational complexity
 - Located in academic/research institutions, but based on public participation
- ### Global megaprojects

 - Very large scale
 - Interdisciplinary
 - Human subjects
 - Inter-state/international
 - Multiple jurisdictions
 - Cross-sector partners
 - Different cultures
- Advancing knowledge to benefit society, but raising multiple issues of concern...



Pittsburgh Health Data Alliance

Carnegie
Mellon
University

University of
Pittsburgh

UPMC

World-class CS/
machine learning + Medical
research + Deep data,
expertise clinical setting,
commercialization

**What roles can libraries
and librarians play in
such endeavours?**

Secondary data analysis

“The health care field
generates an enormous
amount of data every day.

There is a need, and
opportunity, to mine this data
and provide it to the medical
researchers and practitioners
who can put it to work in real
life, to benefit real people.

Many organizations can fulfill
part of this process, but none
of them are equipped to begin
with raw data, develop an idea
and move that idea directly
into a practice setting.”



Defining digital medicine

"The patient is an enormous repository of information that needs to be harvested as a partnership not only in clinical care but in discovery.

It is the only way we will define wellness and its progression to disease, rather than traditional medicine that defines disease and its progression to death."

(Ausiello in Elenko et al., 2015, p. 456)

Embodied information practices





**PRECISION MEDICINE INITIATIVE®
COHORT PROGRAM**



WHAT IS IT?


Precision medicine is a groundbreaking approach to disease prevention and treatment based on people's individual differences in environment, genes and lifestyle.

The Precision Medicine Initiative® Cohort Program will lay the foundation for using this approach in **clinical practice**.

WHAT ARE THE GOALS?

Engage a group of **1 million or more U.S. research participants** who will share biological samples, genetic data and diet/lifestyle information, all linked to their electronic health records. This data will allow researchers to develop more precise treatments for **many diseases and conditions**.

Pioneer a new model of research that emphasizes **engaged research participants, responsible data sharing and privacy protection**.



Research based on the cohort data will:

- Lay **scientific foundation** for precision medicine
- Help identify new ways to **treat and prevent disease**
- Test whether **mobile devices**, such as phones and tablets, can encourage healthy behaviors
- Help develop the **right drug** for the **right person** at the **right dose**

All of Us

Precision Medicine Initiative

- Launched by President Obama in his January 2015 State of the Union address
- Aims to leverage advances in genomics, emerging methods for managing and analyzing large data sets, and health ICTs to accelerate biomedical discoveries
 - while protecting privacy
- Plans to enroll one million or more volunteers and may include children

“committed to engaging multiple sectors and forging strong partnerships with academic and other non-profit researchers, patient groups, and the private sector to capitalize on work already underway”

Participatory Research



Precision Medicine Initiative – **issues...**

“There’s privacy issues. We’ve got to figure out how do we make sure that if I donate my data to this big pool that it’s not going to be misused, that it’s not going to be commercialized in some way that I don’t know about.

And so we’ve got to set up a series of structures that make me confident that if I’m making that contribution to science that I’m not going to end up getting a bunch of spam targeting people who have a particular disease I may have.”

(Obama, 2016, February 25)

Ethical, legal, and social implications?



the **WHITE HOUSE**
PRESIDENT BARACK OBAMA

ALL OF USSM RESEARCH PROGRAM

All of Us Research Program

October 12, 2016

[Scale and Scope](#)

[Participation](#)

[Program Components](#)

[Funding](#)

[FAQ](#)

[Advisory Groups](#)

[Events](#)

[Announcements](#)

[In the News](#)

[Multimedia](#)

PMI Cohort Program announces new name: the All of Us Research Program

The Precision Medicine Initiative® (PMI) Cohort Program will now be called the *All of Us* Research Program and will be the largest health and medical research program on precision medicine. A set of core values is guiding its development and implementation:



- Participation is open to all.
- Participants reflect the rich diversity of the U.S.
- Participants are partners.
- Participants have access to their information.
- Data will be accessed broadly for research purposes.
- Security and privacy will be of highest importance.
- The program will be a catalyst for positive change in research.

The name change reflects these values. We will invite people from all across the U.S. to raise their hands to be one of a million or more participants who will contribute their health information. This information will form the basis of a data set that researchers will be able to analyze to identify better ways to prevent and treat diseases that are based on individual health, environment, and lifestyle.

To truly reflect the far-reaching nature of the program, NIH developed a name that would convey the inclusivity and openness that are hallmarks of PMI research.

Email Updates

Sign up to receive email updates about the Precision Medicine Initiative.

[Sign up for updates](#)

Related Links

[PMI Working Group Final Report](#)

[pdf](#)

[NEJM Perspective: A New Initiative on Precision Medicine](#)

[White House Precision Medicine Web Page](#)

[White House Fact Sheet: President Obama's Precision Medicine Initiative](#)

[Precision Medicine Initiative and Cancer Research](#)

[Precision Medicine Initiative YouTube Channel](#)

Values
statement





About PGP



Harvard PGP is “an open science research project...designed to create **public** scientific resources that everyone can access by bringing together genomic, environmental, and human trait data donated by our participants”

- Founded at Harvard Medical School in 2005, now a Global Network involving Canada (University of Toronto), the UK (UCL) and Austria (Austrian Academy of Sciences)
- Harvard PGP is staffed by a small, largely volunteer group of researchers, engineers, and ethicists who are all pioneers in their fields
- Members of the Global Network follow a common set of guidelines, but the quantity and quality of information on national sites varies significantly

“Privacy, confidentiality and anonymity are impossible to guarantee in a...research study where public sharing of genetic data is an explicit goal”



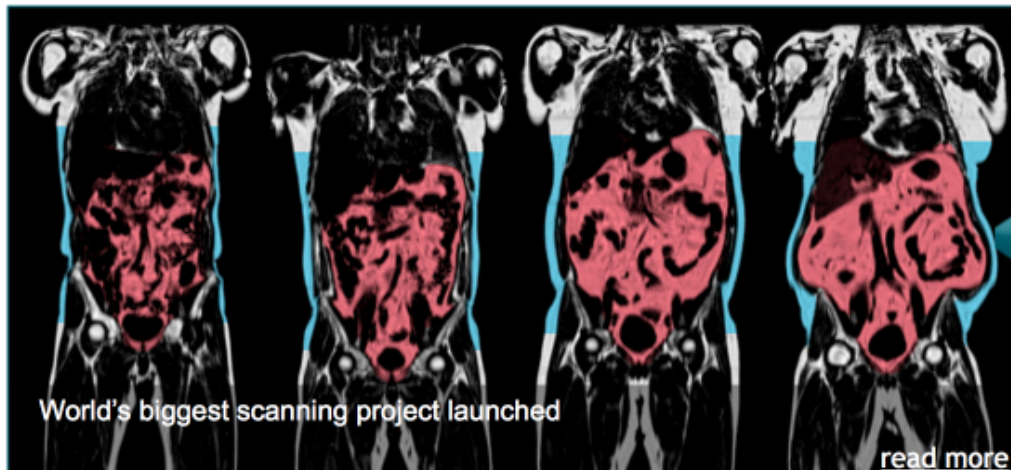
Personal Genome Project



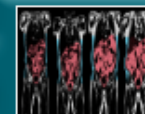
Guidelines of the Global PGP Network

- a) **Public Data.** Participants are invited to share genomic and trait data using a CC0 waiver
- b) **Non-anonymous.** Risks of participant re-identification are addressed upfront as part of the consent and enrollment process
 - **Neither anonymity nor confidentiality of their data is promised to participants**
- c) **Equal access.** Participants are given timely and complete access to their individual data i.e., raw data and not just summary results “where feasible”
- d) **Oversight.** Each member must maintain current Institutional Review Board [Research Ethics] or **local equivalent approval**
- e) **Not for profit.** Managed or sponsored by a non-profit organization (or local equivalent).
 - A member shall not sell or license participant data or tissues “**other than purposes of reasonable cost recovery**”

Pretty Good Privacy?



Annual Meeting
presentations available
now



World's biggest
scanning project
launched



Scanning: looking at the
whole person

Participants

- ☒ [Update your contact details](#)
- ☒ [World's biggest scanning study launched](#)
- ☒ [Find out how the resource is being used](#)
- ☒ [Participant Events](#)

[General Practice linkage →](#)[Data Showcase →](#)[Annual Meeting: watch again](#)

Scientists

- ☒ [Data Showcase](#)
- ☒ [Activity data available in Data Showcase](#)
- ☒ [Video: How to Register and Apply](#)
- ☒ [Annual Meeting: watch again here](#)

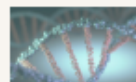
News



Annual Meeting
presentations
available now



World's biggest
scanning project
launched



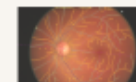
Inherited gene
changes take years off
life expectancy



Active commuting
helps lower body fat
and BMI



Short man or
overweight woman?
Your size could make
you poorer



Assessment
measures provide
insight into common
eye disorders



Background

“A major national health resource”

- Registered charity
- Est. by Wellcome Trust, MRC, Dept. of Health, Scottish Gov., and NW Regional Dev. Agency; funded by Welsh Dev. Agency, BHF, and Diabetes UK)
- Hosted by U. Manchester, supported by NHS
- Open to *bona fide* researchers anywhere in the world, including those funded by academia and industry
- Aims to improve prevention, diagnosis and treatment of life-threatening illnesses
- Recruited 500,000 people aged 40-69 in 2006-2010
- Participants have undergone measures, provided blood, urine and saliva samples, and detailed personal information
 - and agreed to have their health followed

“...to help scientists discover why some people develop particular diseases and others do not”



Best Ethical Practice?

UK Biobank wants to be “a model not only for best science but for best ethical practice too, in relation to these big biobank projects”

Professor Roger Brownsword, Chair (2011-2015)

UK Biobank Ethics and Governance Council (UKEGC)

<http://www.ukbiobank.ac.uk/ethics/>



What are some of the “best science” and “best ethical practice” lessons that can be learned from UK Biobank?



biobank^{uk} “...a precedent-setting case”

- Researchers wanted to use UK Biobank to identify people to invite into a separate study
- They asked UK Biobank to send an introductory email to its participants pointing to the website of the new study
- Offering such a recruitment mechanism could benefit the research community
 - But take time and resources that could be used elsewhere
- In what circumstances would it be acceptable for Biobank to divert resources in this way?
 - How should *ad hoc* third-party re-contacts be accommodated?
- UKBEGC proposed two options
 - Create a dedicated webpage to provide neutral information about (approved) studies
 - Provide a withdrawal category allowing Biobank participants opt-out from email invitations

The project was approved as a pilot subject to fitting with Biobank’s timetable of re-contacts and will be used to draw up a framework for future requests



Participant Centered Consent Toolkit

Sage Bionetworks developed and maintains the *Participant-Centered Consent (PCC) toolkit*.

The goal of the PCC toolkit is to transform the concept of consent from a signature on a legal form to a process that educates, and to allow as many people as possible to engage in consented studies. The PCC toolkit is openly licensed and may be used in any context, by any user, for any purpose, as long as attribution is provided back to Sage Bionetworks.

The PCC toolkit contains a variety of resources for developing a consent process:

- PCC visual dictionary of icons and concept animations
- eConsent workflows
- Design documents and templates
- FAQ

**“...from a signature
on a legal form to a
process that educates”**



Review – Data roles of libraries and librarians

Current roles

- **Access management**
 - facilitating external dataset use
- **Collection building**
 - auditing and appraising data assets, cataloging and archiving datasets
- **Data literacy**
 - educating students and faculty about managing research data
 - preparing graduates for working with data in employment
- **Publishing support**
 - advising on identifying, citing, sharing, licensing and impact
- **Policy development**

Emergent roles

- **Metadata consultancy**
 - providing specialist expertise to support open data sharing
- **Infrastructure development**
 - hosting data hackathons and facilitating data deposit
 - serving as local platforms for open data in smaller communities
- **Data protection**
 - promoting responsible use of personal patron/student data
- **Data literacy**
 - preparing frontline staff to advise on digital privacy or data profiling



Conclusion – The way forward?

- Mainstream data literacy in our teaching and learner support
 - Move data from the periphery to the centre of our practice
- Synthesize diverse perspectives and positions on data literacy
 - Recognize people have multiple roles in different settings
- Give priority to ethical, legal, and social implications (ELSI)
 - Acknowledge our professional responsibility to engage
- Adopt holistic approaches to information instruction
 - Help people use/create data, information, and knowledge in their personal, social, professional, and scholarly lives
- Work proactively across professional and sectoral boundaries
 - Collaborate with public library open data managers



Selected References

Elenko, E. et al.. (2015). Defining digital medicine. *Nature Biotechnology*, 33(5), 456-461. doi:10.1038/nbt.3222.

Inskip, C. (2014, June 16). Information literacy is for life, not just a good degree [Blog post]. Retrieved from <https://www.cilip.org.uk/blog/information-literacy-life-not-just-good-degree>.

Obama, B. (2016, February 25). *Remarks by the President in Precision Medicine Panel Discussion*. Retrieved from <https://obamawhitehouse.archives.gov/the-press-office/2016/02/25/remarks-president-precision-medicine-panel-discussion>.

Pryor, G. (2012, February 6). *Re-skilling for research*. Retrieved from <http://www.dcc.ac.uk/news/re-skilling-research-observations-rluk-report>.

Ruleman, A. B. et al. (2017, March). Show me the learning: Navigating information literacy through multiple life perspectives. Retrieved from <http://www.ala.org/acrl/conferences/acrl2017/papers>.

Any Questions?

Sheila Corrall

Department of Information Culture
and Data Stewardship

School of Computing and Information

scorrall@pitt.edu

www.ischool.pitt.edu/people/corrall.php

