Department of Information Culture and Data Stewardship

Transferability of Data-Related Roles and Competencies

Sheila Corrall
scorrall@pitt.edu

UNIVERSITY OF PITTSBURGH
1787

The Next Frontier – Big Data 2.0

Four Questions for Debate

- What does it mean to have data awareness and understanding in the new world of massive open online data and continuous participatory research programs?
- How can information specialists capitalize on their evolving data-related roles and competencies in the Big Data era?
- Where should librarians concentrate their efforts to create real value for the individuals and communities they serve?
- How can practitioners collaborate to make a big difference in our fast-moving data-rich society?

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

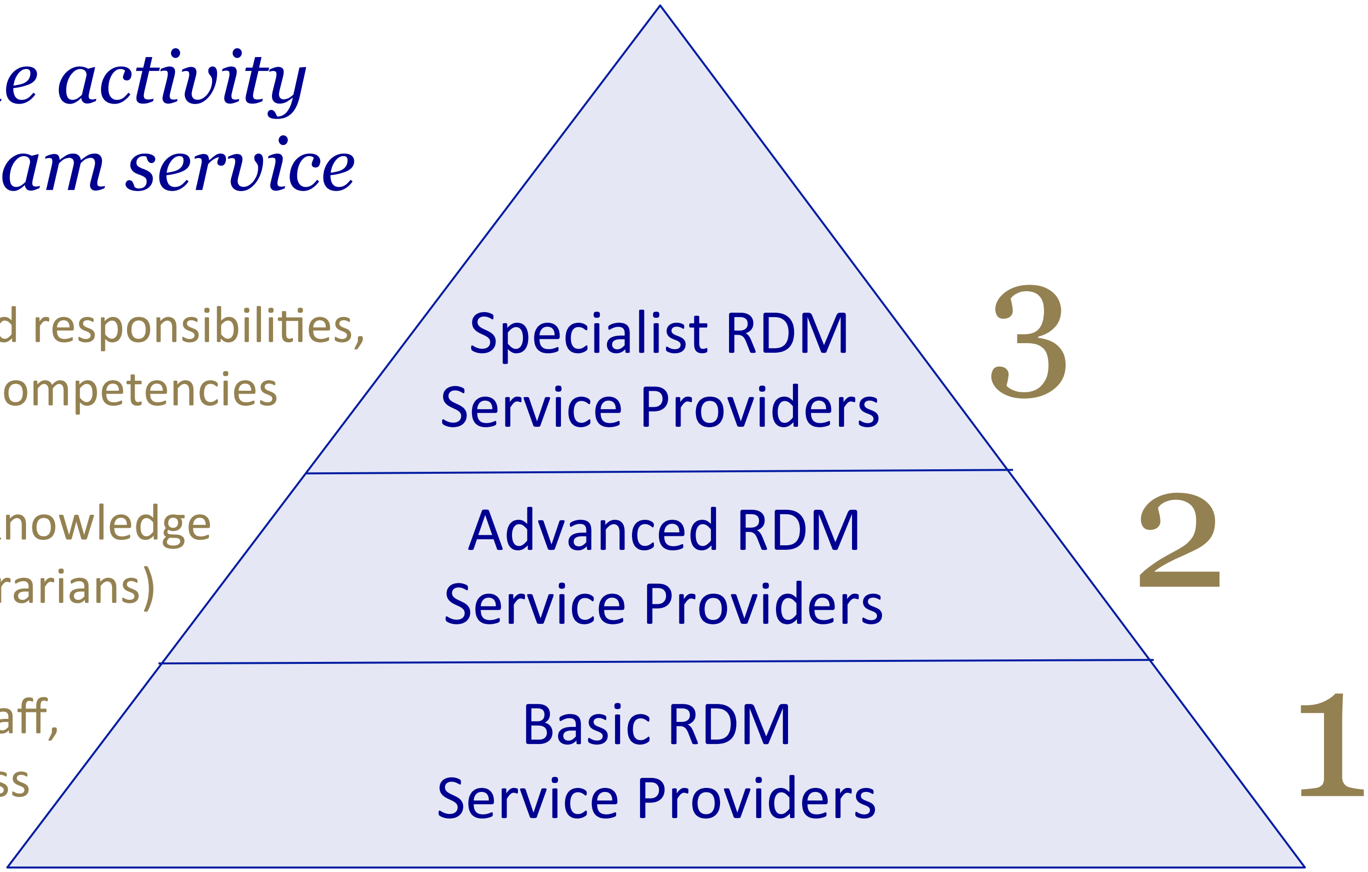
- Data centres/repositories
- Linked data
- Data analytics
- Data warehouses
- Data visualization
- Data journals/papers
- Data citation
- Text and data mining

Beyond Research Support

- 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

RDM Tiered Service Delivery Model

From niche activity to mainstream service



3 Specialist RDM Service Providers

Explicit data-related responsibilities, with in-depth competencies

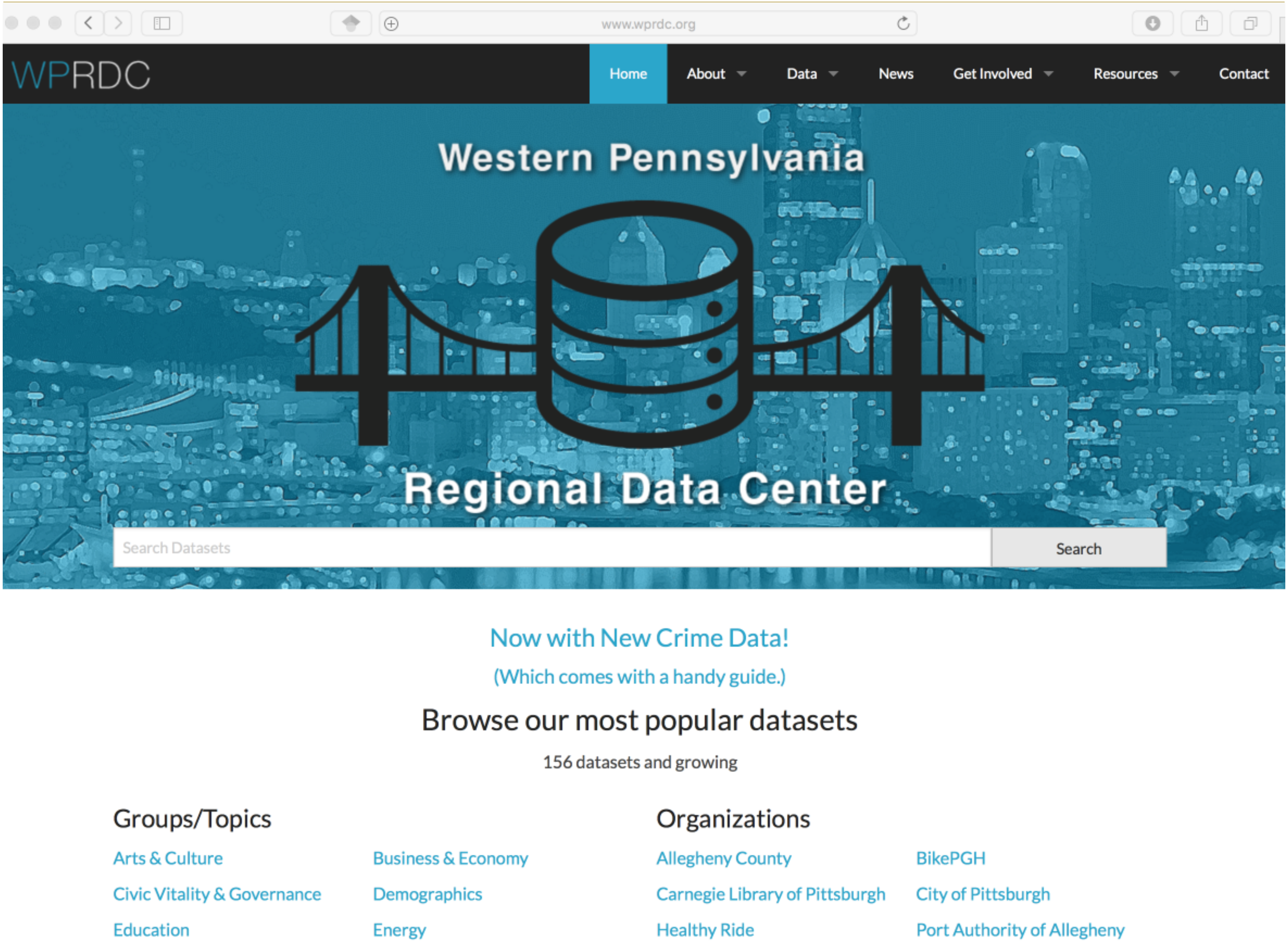
2 Advanced RDM Service Providers

Discipline-based knowledge (e.g., liaison librarians)

1 Basic RDM Service Providers

All public-facing staff, general awareness

(Mattern, Brenner & Lyon, 2016, p. 29)



Now with New Crime Data!

(Which comes with a handy guide.)

Browse our most popular datasets

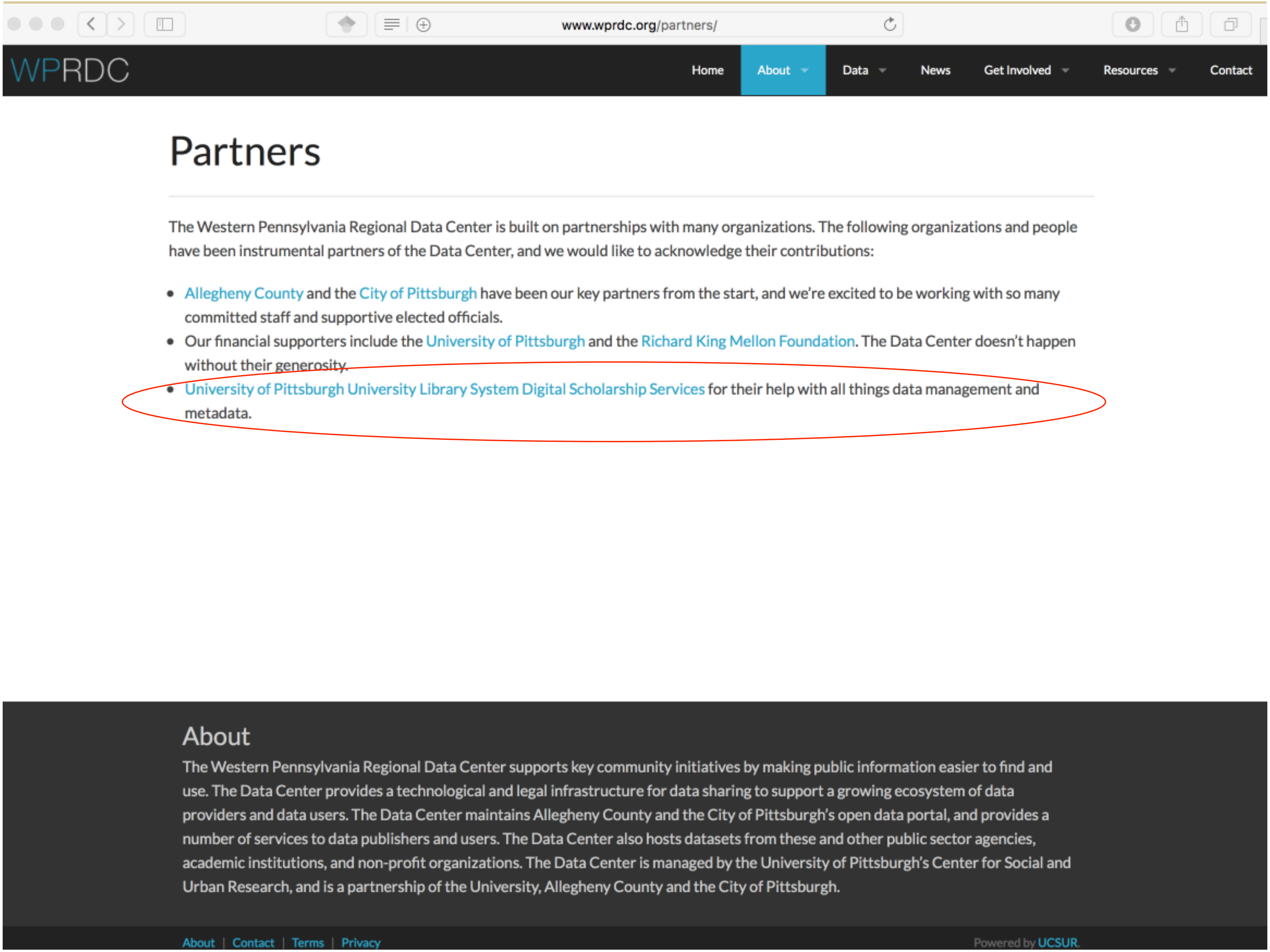
156 datasets and growing

Groups/Topics

- Arts & Culture
- Business & Economy
- Civic Vitality & Governance
- Demographics
- Education
- Energy

Organizations

- Allegheny County
- BikePGH
- Carnegie Library of Pittsburgh
- City of Pittsburgh
- Healthy Ride
- Port Authority of Allegheny



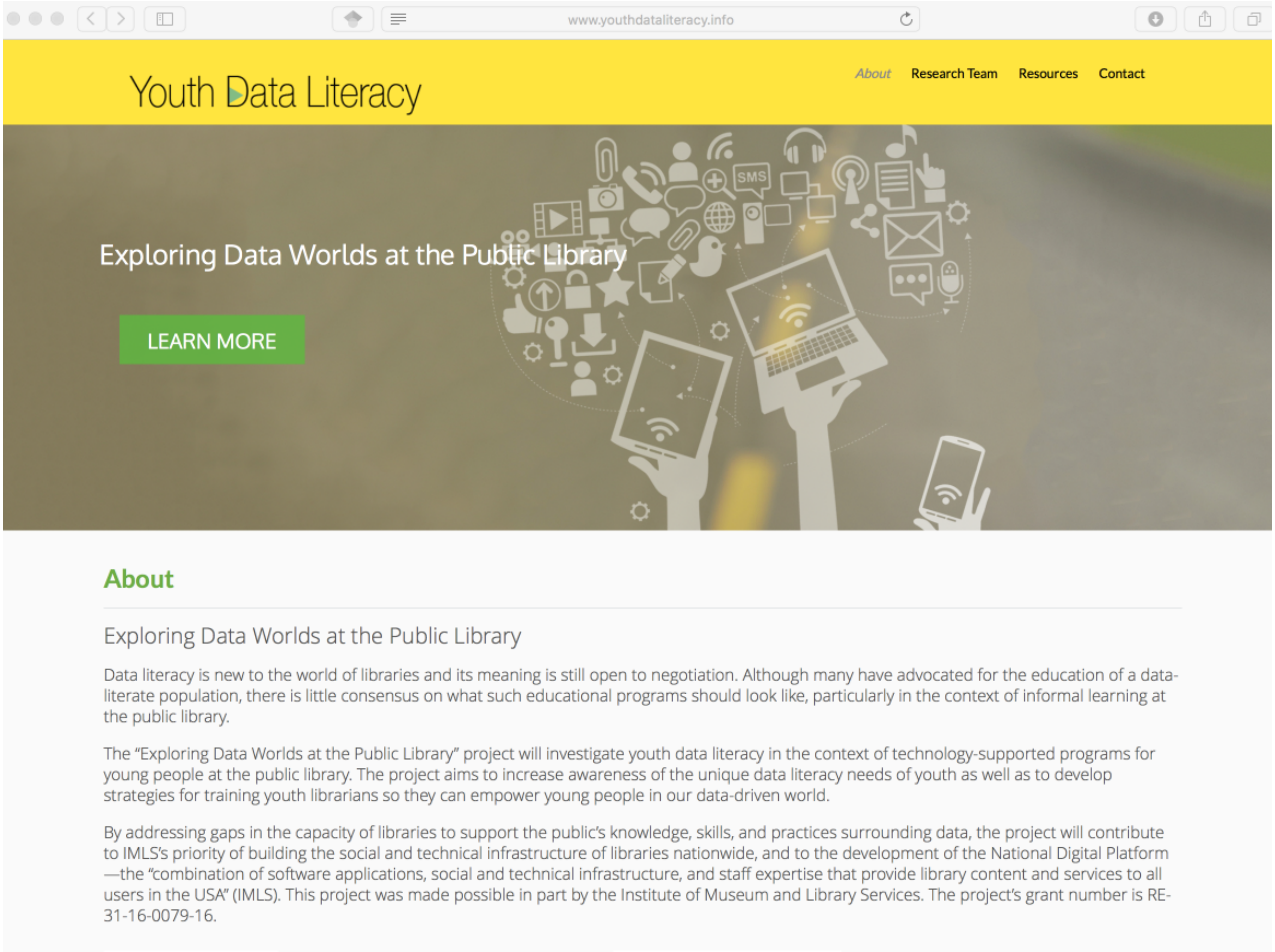
Partners

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.



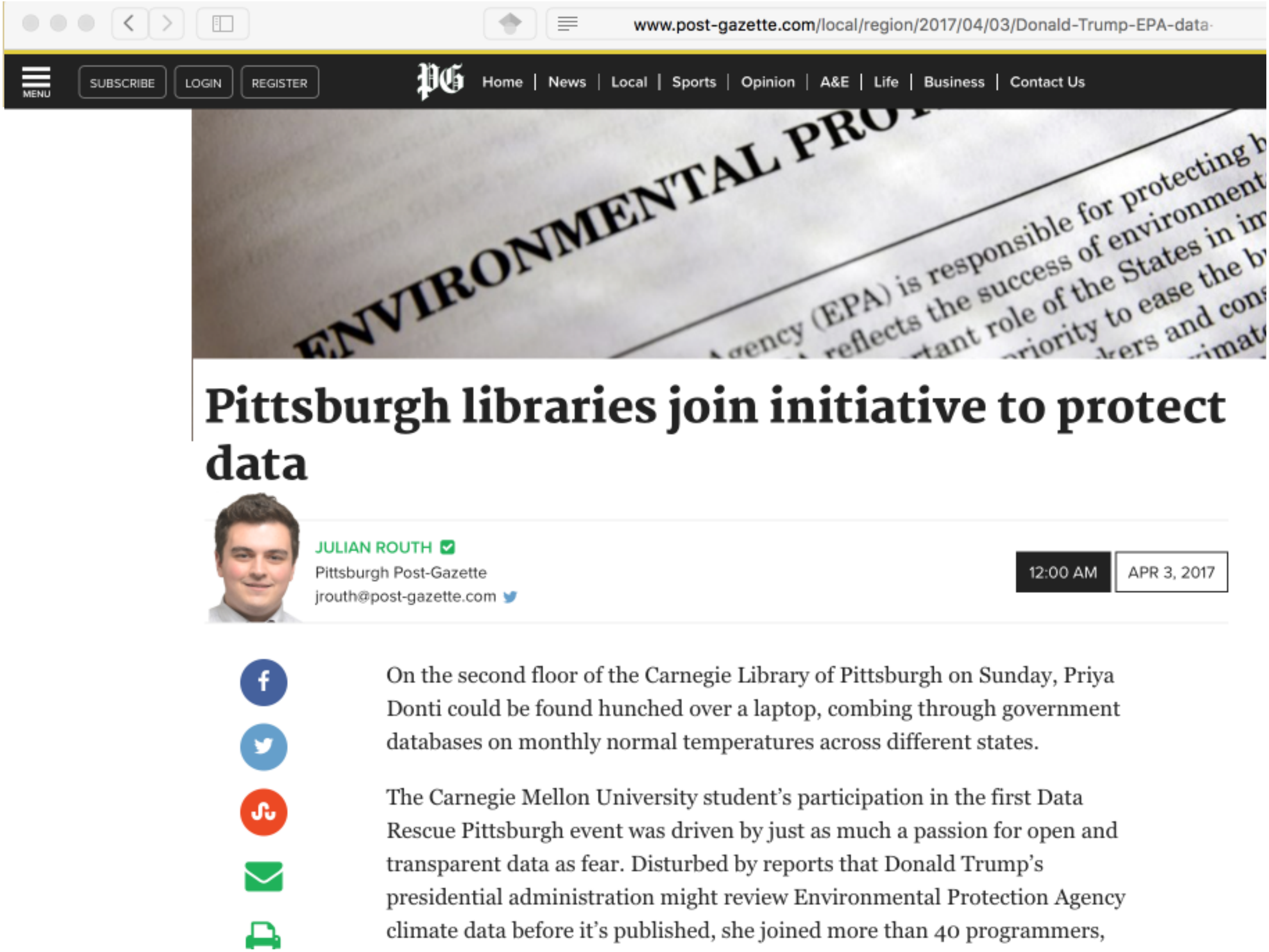
About

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.



Pittsburgh libraries join initiative to protect data

JULIAN ROUTH
Pittsburgh Post-Gazette
jrouth@post-gazette.com

12:00 AM APR 3, 2017

On the second floor of the Carnegie Library of Pittsburgh on Sunday, Priya Donti could be found hunched over a laptop, combing through government databases on monthly normal temperatures across different states.

The Carnegie Mellon University student's participation in the first Data Rescue Pittsburgh event was driven by just as much a passion for open and transparent data as fear. Disturbed by reports that Donald Trump's presidential administration might review Environmental Protection Agency climate data before it's published, she joined more than 40 programmers,

The Next Frontier – Big Data 2.0

- Converging e-science, business intelligence, crowdsourcing, big data analytics, social media and Web 2.0 technologies
- Enabling broader and deeper applications of analytical tools
- Located in academic/research institutions, usually based on public participation and often involving commercial players
- Taking very large-scale data-intensive research to new levels of organizational and socio-technical complexity
- Raising ethical, legal and policy issues

Massive **O**pen **O**nline **D**ata **S**tudies

MOODS

Pittsburgh Health Data Alliance

Carnegie Mellon University | University of Pittsburgh | UPMC

World-class CS/ machine learning + Medical research expertise + Deep data, 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."

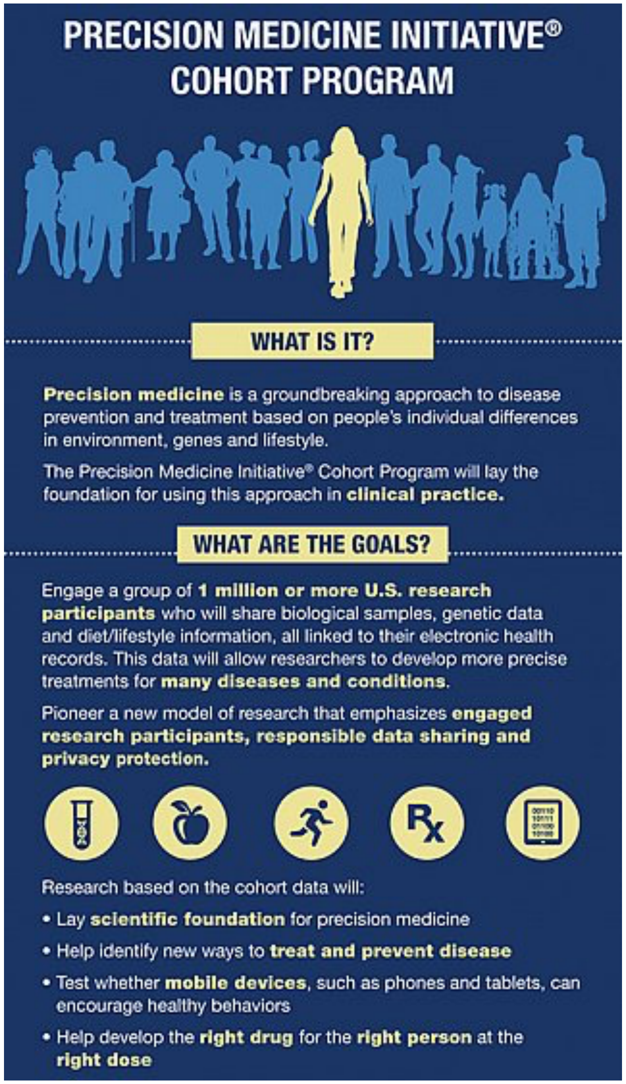
Defining digital medicine

> “**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



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

- 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

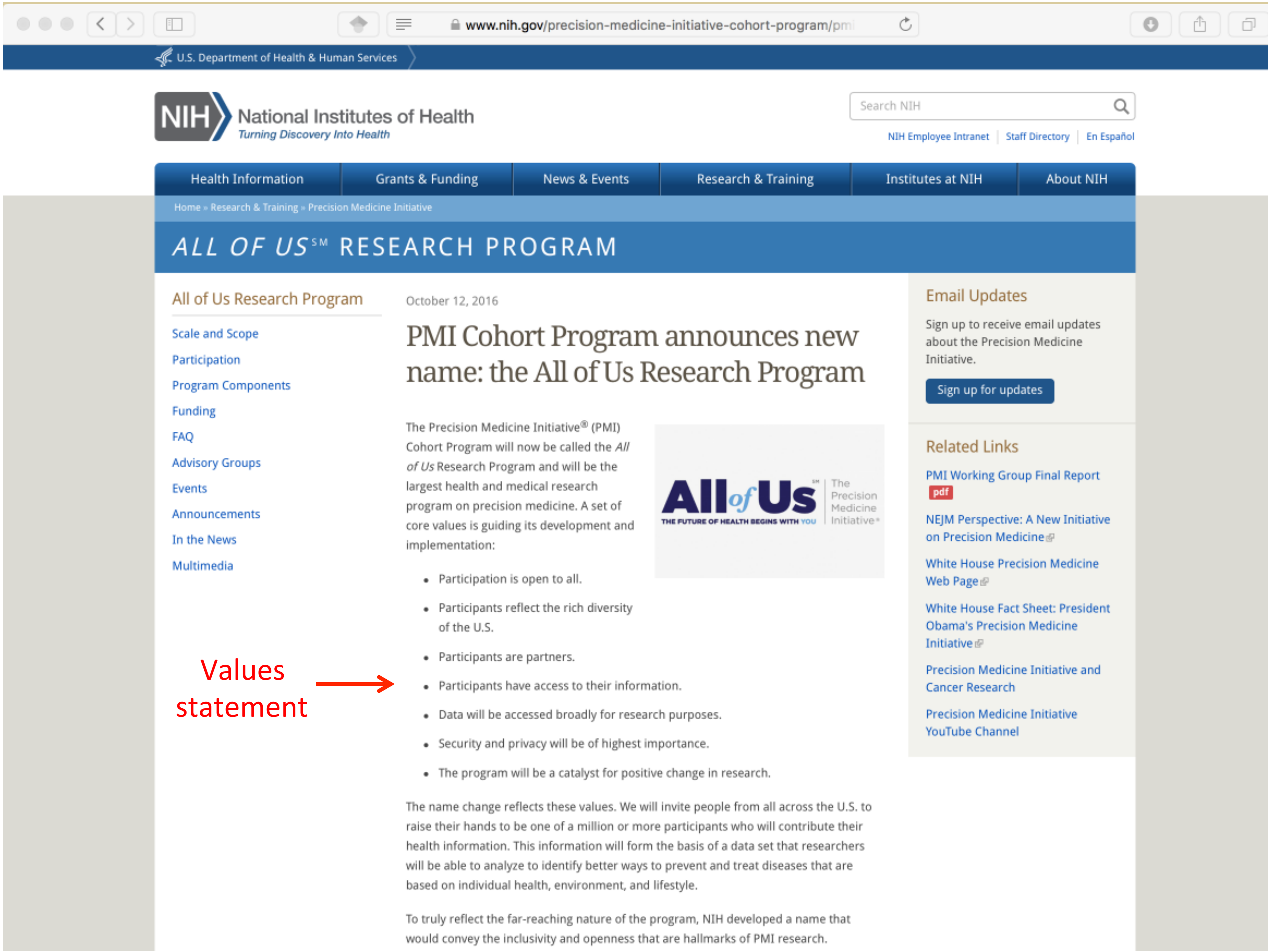
Precision Medicine Initiative – **issues…**

> “**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 US℠ RESEARCH PROGRAM

October 12, 2016

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.

Related Links

PMI Working Group Final Report

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

genom austria

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 except for purposes of **"reasonable cost recovery"**

Pretty Good Privacy?

Call Us On: 0800 0 276 276

Your feedback is important to us, tell us what you think

World's biggest scanning project launched

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 biobank uk

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/

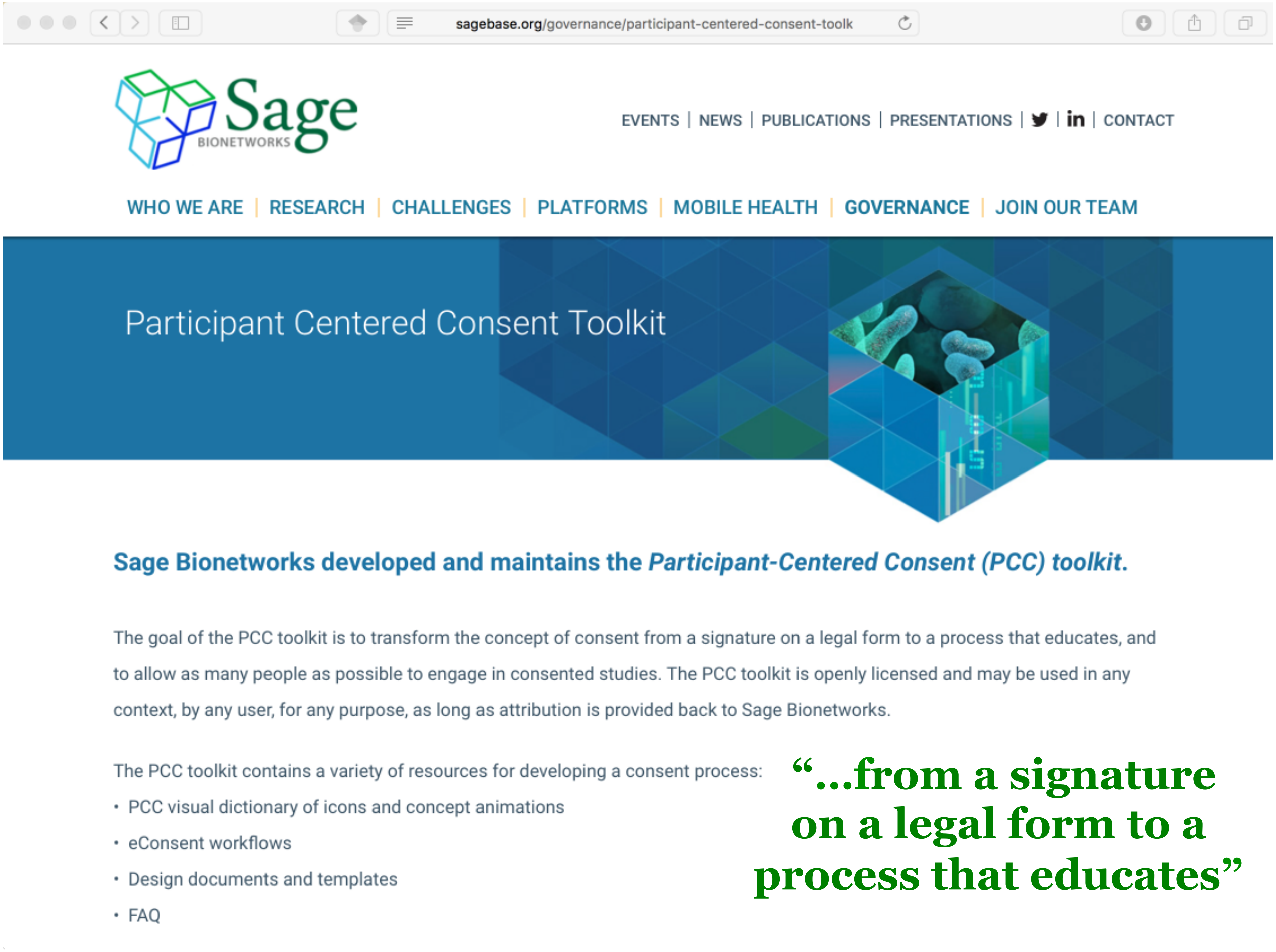
So, what are the ‘best science’ and ‘best ethical practice’ lessons to be learned from UK Biobank?

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

UK BIOBANK ETHICS AND GOVERNANCE COUNCIL ANNUAL REVIEW 2015



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”

Issues arising from **Big Data 2.0** projects

Legal compliance

- Privacy laws
- Data protection legislation
- Right to be forgotten
- Genetic information laws
- Freedom of information
- Intellectual property

  e.g., patenting human genes (cf. EU and US case ruling)
- Licensing/contractual issues
- Publishing

Ethical challenges

- Privacy
- Anonymity
  - protection from bad actors e.g., cybercriminals, hactivists
- Monetization
  - selling of health data
- Conflicts of interest
- Informed consent
- Solicitation of donors for participation in other studies

Policy questions arising from Big Data 2.0

- How and by whom will health data/big data be **preserved** and made retrievable for and by future stakeholders?
- What guidelines and requirements are needed for **publishing** related to health data/big data?
- Who needs to have a voice in policy-setting and policy-making, and who should craft the governing **policies** and codes of ethics?
  - ☞ Given the pace of change, how often should policies and codes be reviewed and updated?
- What oversight and enforcement mechanisms are needed to ensure **compliance**?
  - ☞ What are the penalties for piracy of health data or malfeasance, negligence, willful blindness, and harmful impacts on human subjects?
  - ☞ What protections are available or need to be developed and codified for whistleblowers who report lapses and breaches of compliance?

Big Data 2.0 – Potential roles for info pros

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…

- **Data ethics** – monitoring practices and advocating or contributing to ***policy*** for data protection and research integrity
- **Data literacy** – extending education to cover personal, social, professional and scholarly ***contexts*** of data creation, sharing and use
- **Digital curation** – applying repository and RDM know-how (e.g., ***metadata*** advice and consultancy)
- **Interdisciplinary facilitators** – helping multidisciplinary teams navigate unfamiliar ***territory***

Conclusion – Critical roles for LIS

- Mainstream data literacy and adopt holistic approaches to include data handling in education, work, and everyday life
- Raise awareness of ethical, legal, and social implications (ELSI) of large-scale participatory data-intensive projects
- Work proactively across professional and sectoral boundaries to share and transfer essential know-how (e.g., metadata)

Provide a human-centred perspective – The conscience of the big data world

Acknowledgement – Special thanks to my collaborator, Dr. James D. (Kip) Currier, for his expert analysis of the ethical, legal and policy issues arising from the Big Data 2.0 case studies.