



Building a Global Data Community

Dr. Francine Berman

Chair, Research Data Alliance / US

Hamilton Distinguished Professor in
Computer Science, Rensselaer
Polytechnic Institute

Some Background

Professional

- First job (Asst. Professor) at Purdue University
- HPC Endowed Chair at UC San Diego
- Director, San Diego Supercomputer Center
- Vice President for Research at Rensselaer
- Hamilton Distinguished Professor at Rensselaer and Chair, Research Data Alliance/U.S.

Personal

- Two great children: Emily (27) – choreographer and Dancer, Nicholas (24) – Ph.D. student in Math
- Supportive partner and husband Mark
- Do what's important and never take your biosketch too seriously



It's a Digital World



Physical
Infrastructure



Entertainment



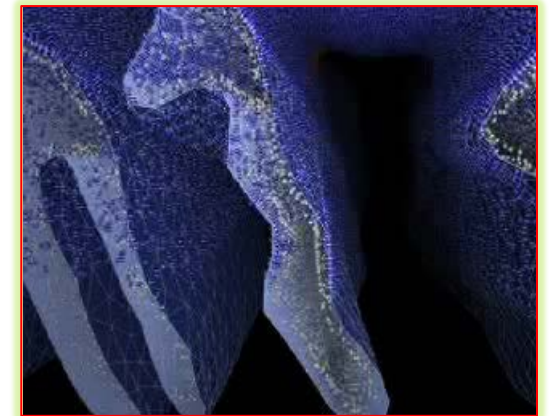
Commerce



Health



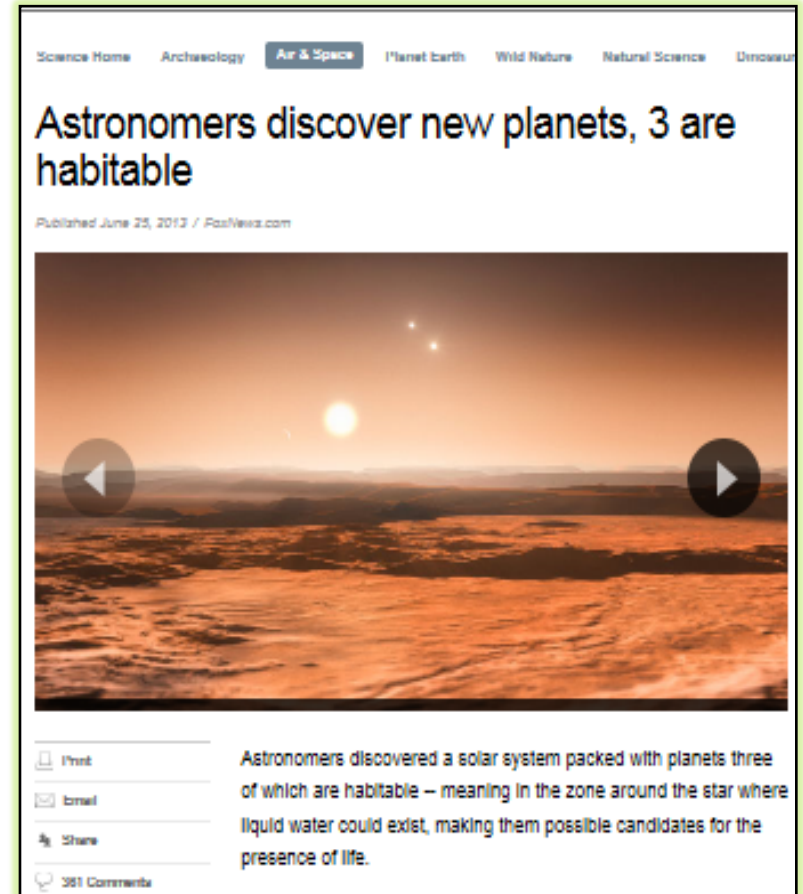
Communication /
Community



Research

Data-Driven Research: More life in the universe?

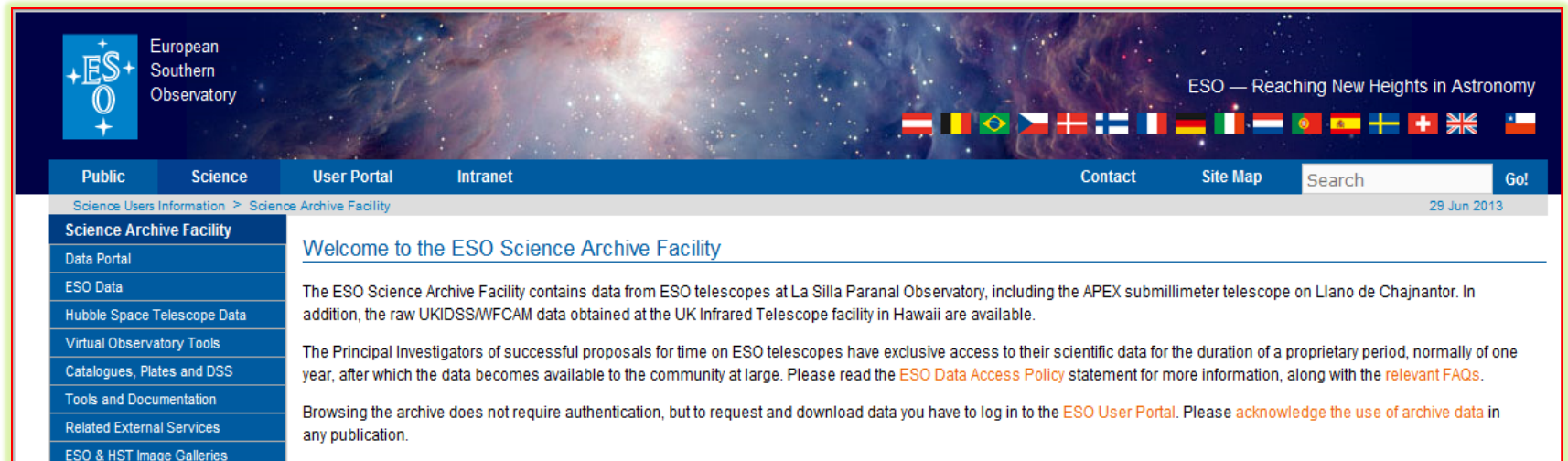
- June, 2013: **Announcement of discovery of 3 “habitable” planets around the star Gliese 667C** -- ~129 trillion miles (~ 22 light years) away
- Gliese 667C is 1/3 the size of the sun. Habitable (liquid water could exist, etc.) planets are more massive than earth.
- International research team includes researchers from Germany, U.K., U.S., Finland, Chile



Data-Driven Discovery

“We identified 3 strong signals in the star before, but it was possible that smaller planets were hidden in the data. **We re-examined the existing data, added some new observations and applied two different data analysis methods** especially designed to deal with multi-planet signal detection.”

Guillem Anglada-Escude'
University of Gottingen, Germany



European Southern Observatory

ESO — Reaching New Heights in Astronomy

Public Science User Portal Intranet Contact Site Map Search Go!

Science Users Information > Science Archive Facility 29 Jun 2013

Science Archive Facility

- Data Portal
- ESO Data
- Hubble Space Telescope Data
- Virtual Observatory Tools
- Catalogues, Plates and DSS
- Tools and Documentation
- Related External Services
- ESO & HST Image Galleries

Welcome to the ESO Science Archive Facility

The ESO Science Archive Facility contains data from ESO telescopes at La Silla Paranal Observatory, including the APEX submillimeter telescope on Llano de Chajnantor. In addition, the raw UKIDSS/WFCAM data obtained at the UK Infrared Telescope facility in Hawaii are available.

The Principal Investigators of successful proposals for time on ESO telescopes have exclusive access to their scientific data for the duration of a proprietary period, normally of one year, after which the data becomes available to the community at large. Please read the [ESO Data Access Policy](#) statement for more information, along with the [relevant FAQs](#).

Browsing the archive does not require authentication, but to request and download data you have to log in to the [ESO User Portal](#). Please [acknowledge the use of archive data](#) in any publication.

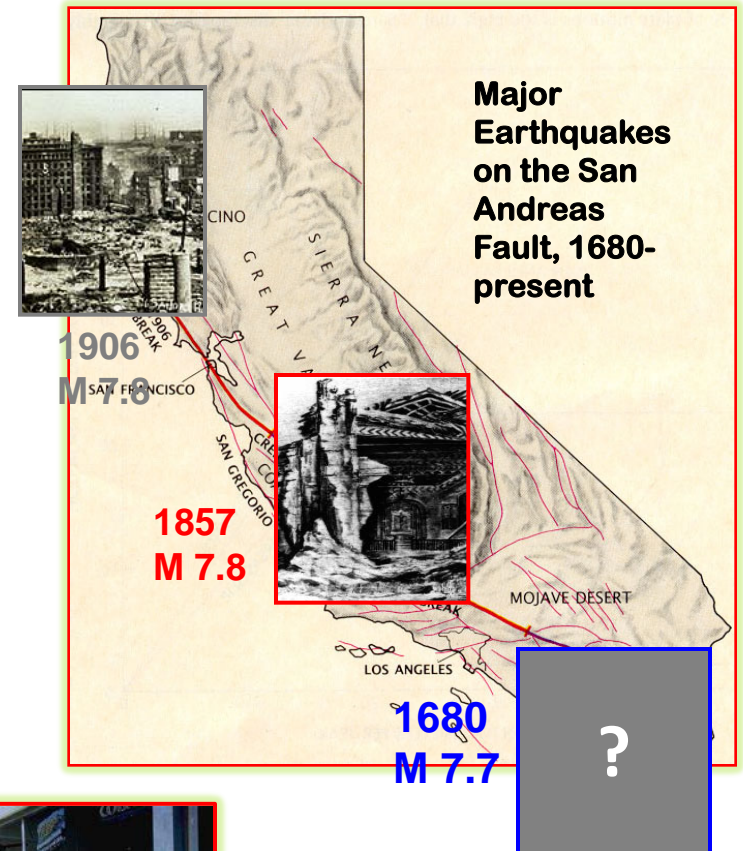
The Data “Backstory”

- Data mostly from ESO Science Archive Facility. Data sets include HARPS-TERRA Doppler measurements, HIRES, and PFS.
- Observations based in part from W. M. Keck Observatory, Magellan team (Doppler measurements), SIMBAD Data base (CDS in France).
- **Data infrastructure needed:**
 - Data management, hosting and preservation infrastructure
 - Data analysis tools
 - Standards for astronomy data and metadata. Organizational policy and adoption in the use of standards.
 - Community practice and technological infrastructure enabling data sharing between researchers



Data-Driven Research – Earthquake Simulation

- Earthquake simulations enable
 - Estimation of seismic risk
 - Emergency preparation, response and planning
 - Design of next generation of earthquake-resistant structures
- Simulations combine large-scale data collections, high-resolution models, supercomputer runs



6.7 M earthquake in Northridge California 1964, earthquake brought estimated \$20B damage

Earthquake Simulation

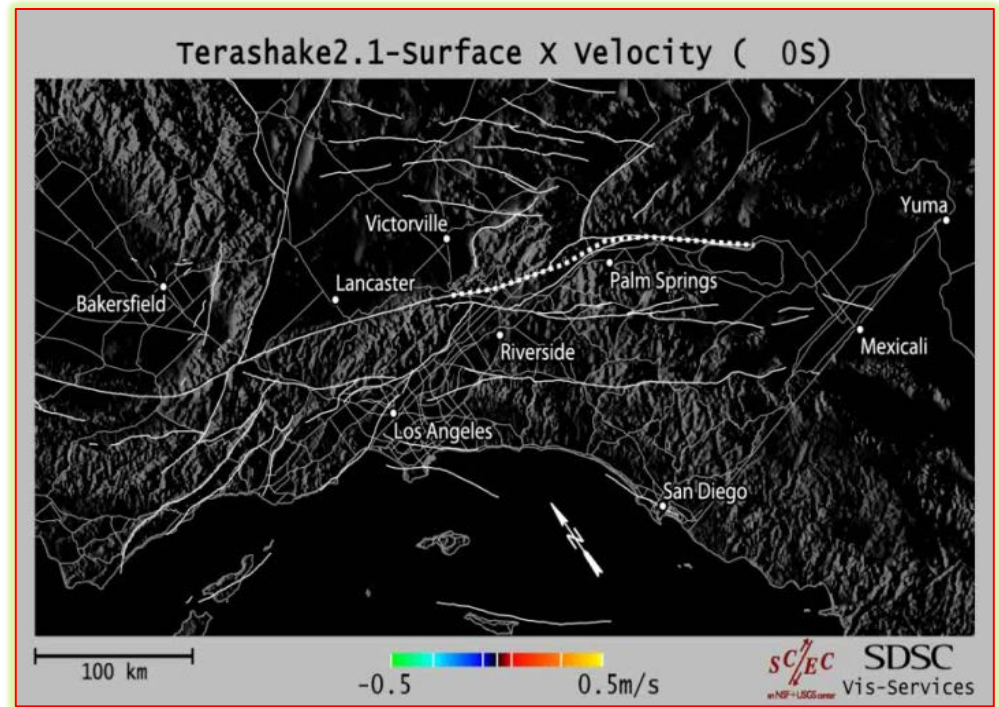
- **Input for supercomputer simulation:**

- 7.7 earthquake model of lower San Andreas fault
- 10 years of sensor data on southern California terrain

- **Post simulation**

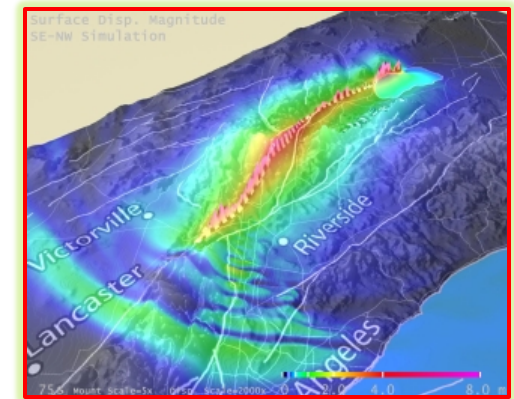
- Additional computation (80,000+ CPU hours) used for visualization of seismic wave propagation and analysis
- Derived data products (velocity magnitude, displacement vector field, cumulative peak maps, statistics, etc.) included in SCEC digital library (>168 TB).

- **Results used for better science and a safer society (better building codes, more informed disaster response planning)**



Better Prediction Accuracy Involves “Bigger” Data

Estimated figures for simulated 240 second period, 100 hour run-time	TeraShake domain (600x300x80 km ³)	PetaShake domain (800x400x100 km ³)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
Time steps	20,000 (.012 sec/step)	160,000 (.0015 sec/step)
Surface data	1.1 TB	1.2 PB
Volume data	43 TB	4.9 PB



Technical Infrastructure for Data-Driven Research



The Digital Research Data Life Cycle: Data from birth to death / immortality

Create

Data creation / capture / gathering from

- laboratory experiments
- fieldwork
- surveys
- devices
- simulation output ...

Edit

- Organize
- Annotate
- Clean
- Filter



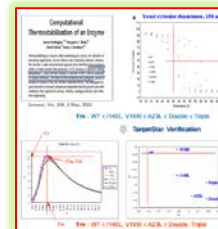
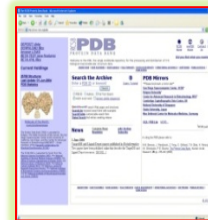
Use / Reuse

- Analyze
- Mine
- Model
- Derive additional data
- Visualize
- Input to instruments / computers / devices



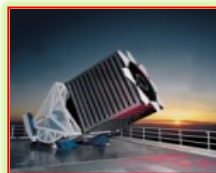
Publish,
Disseminate

- Disseminate
- Create portals / data collections / databases
- Couple with literature

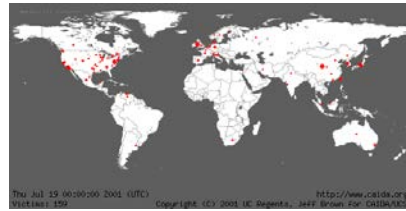
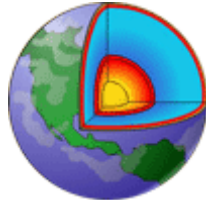
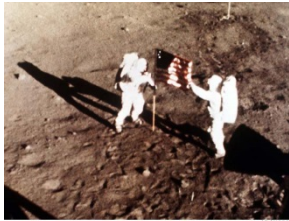


Preserve /
Destroy

- Store / preserve
- Store / replicate / preserve
- Store / ignore
- Destroy



Technical Infrastructure Needed to Scaffold Data-Driven Research



Data
Access

Data
Sharing

Data
Visualization

Data
Analysis

Data
Services

Data
Mining Algorithms

Data
Security

Data
Management

Digital Object
Identifiers

Common
Metadata Standards

Data
Registries

Semantic
Ontologies

Tools and infrastructure
that promote Discoverability

Data
Preservation

Data
Storage

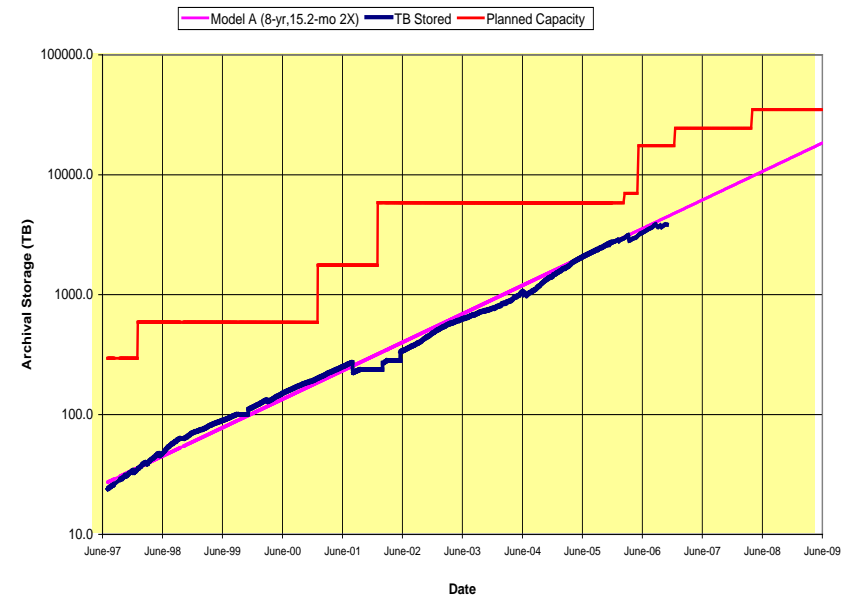
Data Citation
Standards

Research Data Centers Require Viable Economic Support

Costs include

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, policy, etc. ...

Resources and Resource Refresh



SDSC Data Storage Growth '97-'09

- Most valuable data replicated
- As research collections increase, storage capacity must stay ahead of demand

Who Pays the Data Bill for Open Access Research Data?

It's not

Data

usage

requ

perc

Great

extre

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren *JPH*
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are a source of progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open wheat data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets, services related to curation, preservation, analysis, and visualization. Policies that mobilize publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of

Long-lived
data

The New York Times

The Opinion Pages

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

WE'RE READY TO WORK FOR YOU.

EDITORIAL

We Paid for the Research, So Let's See It

Published: February 25, 2013

The Obama administration is right to direct federal agencies to make public, without charge, all scientific papers reporting on research financed by the government. In a memorandum issued on Friday, John Holdren, the president's science adviser, directed federal agencies with more than \$100 million in annual research and development expenditures to develop plans for making the published results of almost all the research freely available to everyone within one year of publication.

Connect With Us on Twitter
For Op-Ed, follow @nytopinion and to hear from the editorial page editor, Andrew Rosenthal, follow @andyrNYT.



The agencies must submit plans to the [White House Office of Science and Technology Policy](#) within the next six months that will apply to both peer-reviewed scientific papers and digital manuscripts and supporting data.

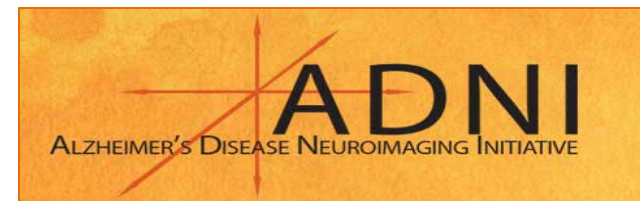
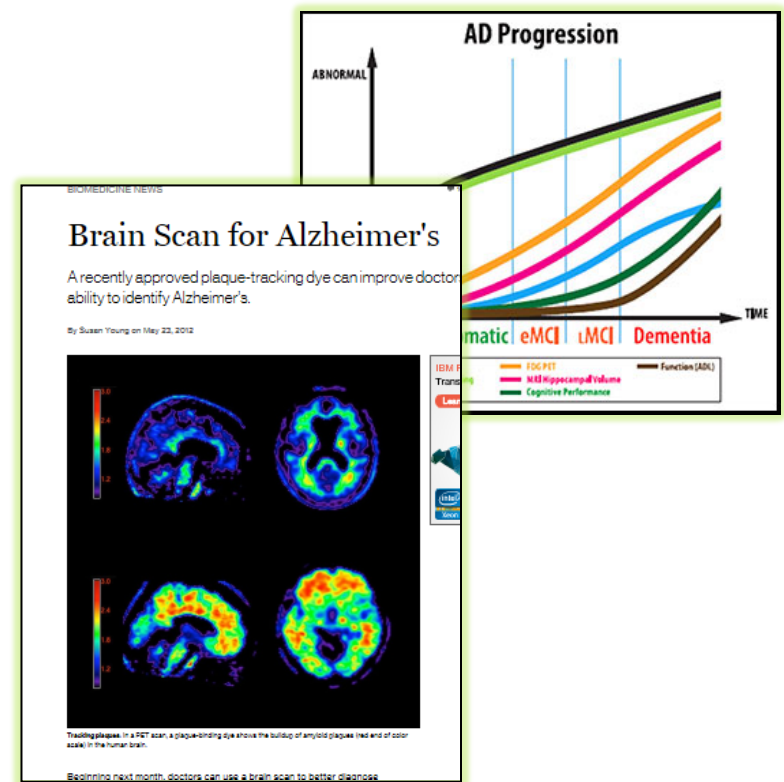
Under current procedures, much of the federally financed research is published in scientific and medical journals that can cost thousands of dollars a

Social Infrastructure for Data-Driven Research

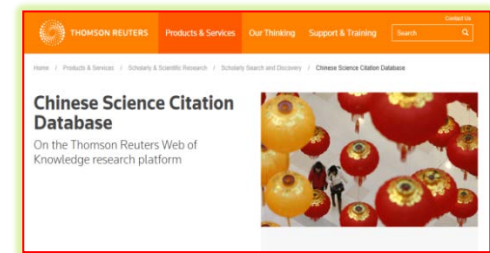
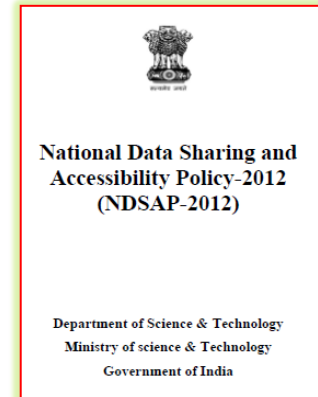


“Social Infrastructure” critical for Data-Driven Research

- Viable Economic Models
- Policy and Governance
- Common Practice
- Organizational Support
- Legal and Regulatory Frameworks
- Science Community Cultures



Technical and Social Data Infrastructure Needed World-wide

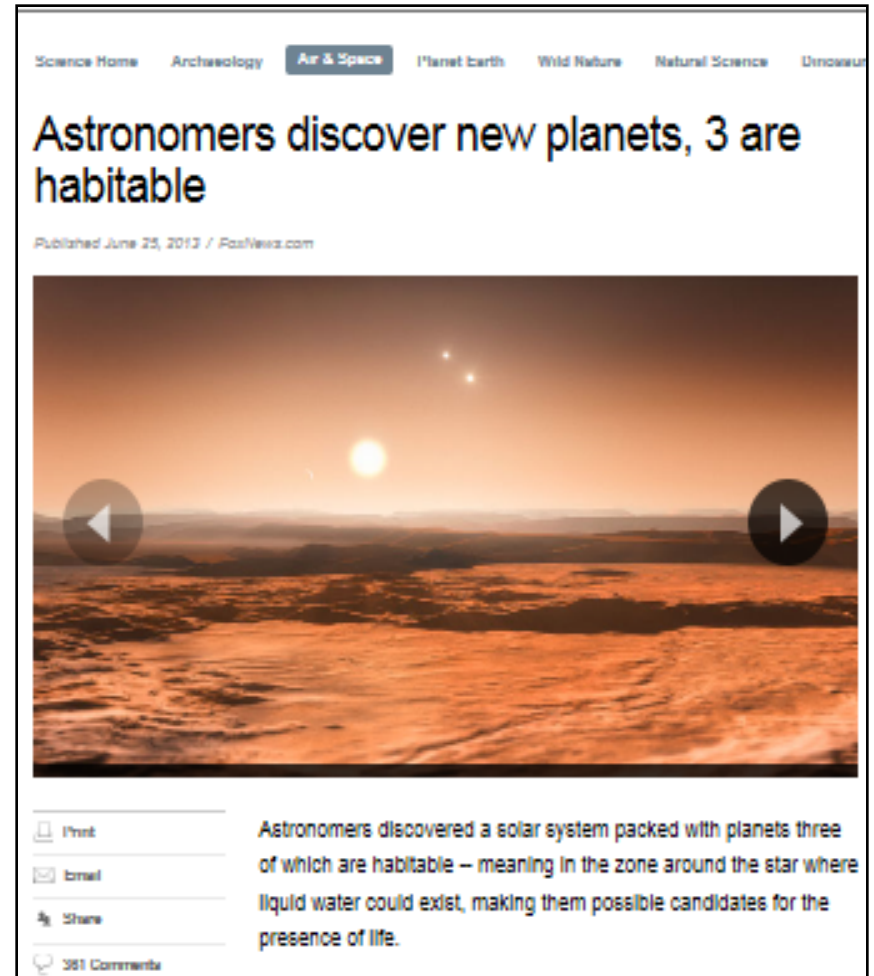


Goal: Research Without Barriers

International research team includes researchers from Germany, U.K., U.S., Finland, Chile

Trans-national data infrastructure:

- **Data management, hosting and preservation** infrastructure
- **Data analysis** tools
- **Standards** for astronomy data and metadata. Organizational **policy** and adoption in the use of standards.
- **Community practice** and technological infrastructure enabling data sharing between researchers



The Research Data Alliance (RDA)

- Global community-driven organization launched in March 2013 to accelerate data-driven innovation
- RDA focus is on building the **social, organizational and technical infrastructure** to
 - reduce barriers to data sharing and exchange
 - accelerate the development of coordinated global data infrastructure



Create → Adopt → Use

RDA members come together to build and use data sharing infrastructure

- Focused pieces of adopted code, policy, infrastructure, standards, or best practices that enable data sharing
- “Harvestable” efforts for which 12-18 months of work can eliminate a roadblock
- Efforts that have substantive applicability to groups within the data community, but may not apply to everyone
- Efforts for which working scientists and researchers can start today

RDA takes an agile approach:

Targeted infrastructure efforts undertaken to drive tangible progress for key communities

- Common metadata standards
- Interoperability / integration framework
- Data access and preservation policy and practice
- Harmonized standards
- Common economic model for sustaining data
- Digital object identifiers
- Tools for data discoverability, etc.

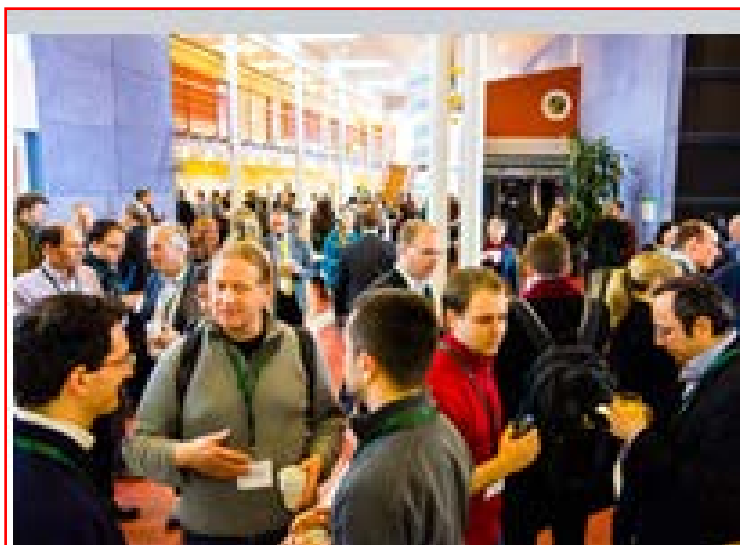


Harmonized standards

RDA Launch March 2013

Gothenburg, Sweden

- Over 200 participants
- 31 countries
- 5 continents
- > 6,400 tweets
- Public, private, academic sectors
- High-profile Govt. and Science speakers




 **najla**
@najlaaa
Jahanian #NSF: the #RDALaunch promises to accelerate pace of scientific discoveries. Data is new currency for science

4 DAYS AGO · REPLY · RETWEET · FAVORITE

 **ResearchDataAlliance**
@resdatall
Another theme emerging: RDA as a "neutral" space across domains, organizations, nations, etc. #RDALaunch

2 DAYS AGO · REPLY · RETWEET · FAVORITE

 **ResearchDataAlliance**
@resdatall
South Africa can't afford not to participate in RDA #RDALaunch

4 DAYS AGO · REPLY · RETWEET · FAVORITE

RDA Interest Groups focusing on infrastructure needed for data sharing and exchange

Current RDA Interest Groups

- Agricultural Data Interoperability
- Big Data Analytics
- Brokering
- Data in Context
- Defining Urban Data Exchange for Science
- Engagement Group
- Legal Interoperability (joint with CODATA)
- Marine Data Harmonization
- Data Citation
- Preservation e-Infrastructure
- Publishing Data
- Repository Audit and Certification
- Structural Biology
- Community Capability Model
- UPC Code for Data
- Digital Practices in History and Ethnography
- ...

RDA Digital History and Ethnography Interest Group

- What factors increase your risk of getting asthma?
- What factors increase your ability to get better?
- How do organizations respond to asthma?
- How do communities respond to asthma?



Asthma as a socio-cultural-health issue

- How is asthma contracted, experienced and cared for in neighborhoods, cities, and countries?
- **Relevant data:** Health, environmental, population, socio-cultural, historical data, etc. in the form of images, video, oral interviews, surveys, HIPAA-compliant data, etc.



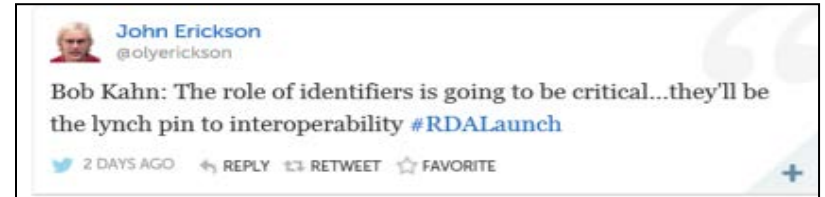
RDA Working Groups Creating a Pipeline of Impact-focused Deliverables

RDA Working Groups

- PID Information Types
- Data Type Registries
- Data Foundation and Terminology
- Practical Policy (pending)
- Metadata Standards (pending)
- Standardization of Data Categories



RDA Persistent Identifier (PID) Information Types Working Group



- **RDA Working Group Focus:**

- Harmonization of basic information types associated with persistent identifiers
- Agreement about the information associated with PIDs allows programmers to implement the same API independent of the PID type being used

- **Impact:**

- Facilitates type harmonization and interoperability between data sharing tools in different infrastructures and domains

- **Deliverables:**

1. Community-vetted list of common information types; framework to introduce more types; mechanisms to develop profiles, collections and typed references
2. Prototype API service for requesting PID information

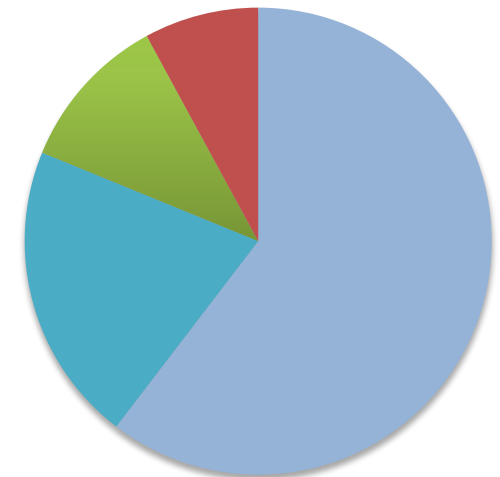
- **Adopters:**

- Infrastructure supported by **CNRI, California Digital Library**
- API will be used in practice by **DKRZ, Data Conservancy, DARIAH** (research data infrastructure in arts and humanities), **EUDAT** (EU Data Infrastructure collaborative)

Current Status: RDA Community = > 700 participants from 44+ countries

Albania	Greece	Portugal
Australia	Iceland	Russia
Austria	India	Serbia
Bangladesh	Iran	Singapore
Belgium	Ireland	South Africa
Bulgaria	Italy	South Korea
Brazil	Japan	Spain
Canada	Krygrystan	Sweden
China	Kuwait	Switzerland
Congo	Netherlands	Taiwan
Czech Republic	New Zealand	Turkey
Denmark	Norway	United Arab Emirates
Estonia	Palestine	United Kingdom
Finland	Poland	United States
France		
Germany		

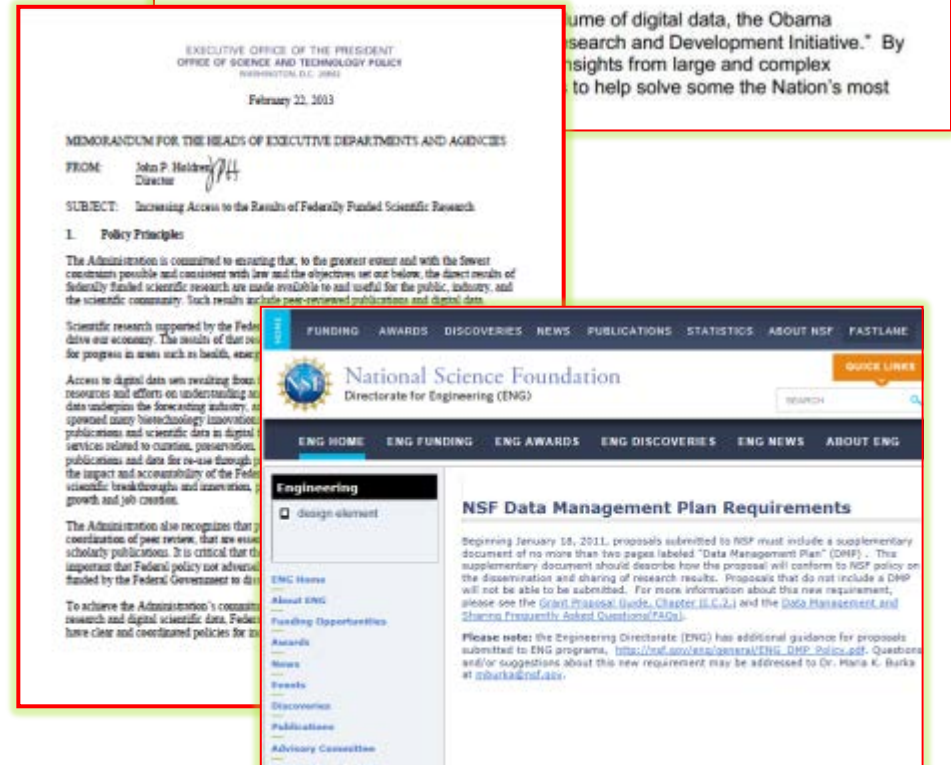
RDA by Sector



■ Academics (61%)
■ Private Sector (21%)
■ Public Sector (11%)
■ Unknown (8%)

RDA and Publicly Funded Research

- **Nationally:** Many countries looking to accelerate data-driven research. **RDA provides a vehicle for implementing the underlying infrastructure required to make new policy approaches work.**
- **Internationally:** RDA provides regional leadership opportunities and promotes global competitiveness and innovation



RDA – How to Get Involved (rd-alliance.org)

- **Join RDA to participate as an individual member.** Register at rd-alliance.org. Membership is free.
- **Join as an Organizational Member** (nominal fee) **or an Organizational Affiliate** (jointly sponsored efforts)
- **Initiate or Join an Interest Group** Community members exploring infrastructure in topical areas
- **Propose or Join a Working Group** (focused 12-18 month efforts with measurable outcomes that accelerate data sharing and exchange)
- **Attend the RDA Plenaries.** Go to rd-alliance.org to register.



RDA Plenary 2: September 16-18 in DC

Please join us for RDA Plenary 2

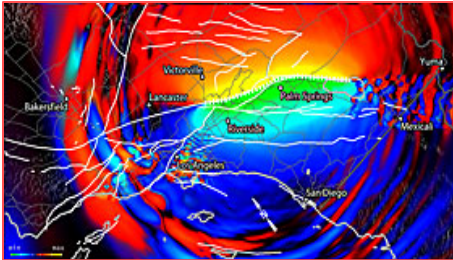
- National Academies and Washington Marriott, September 16-18
- Working Meeting for RDA Interest Groups and Working Groups
- Plenary Speakers and Panels
- RDA Business Meeting and Community Events
- “Neutral space” to convene communities



RDA Plenary 3
in Dublin Ireland,
March 26-28
2014, hosted by
Australia and
Ireland



Building a Global Data Community



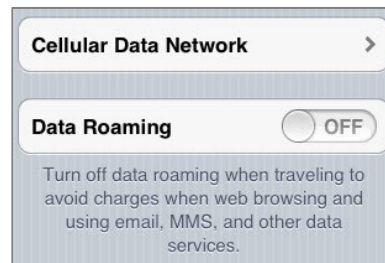
**Technical
Infrastructure**



**Social
Infrastructure**



**Broad
Coordination**



Viable Economic Models



**Policy and
Practice**

Thank You

