RESEARCH DATA ALLIANCE

The Research Data Alliance: Building community and infrastructure for data sharing world-wide

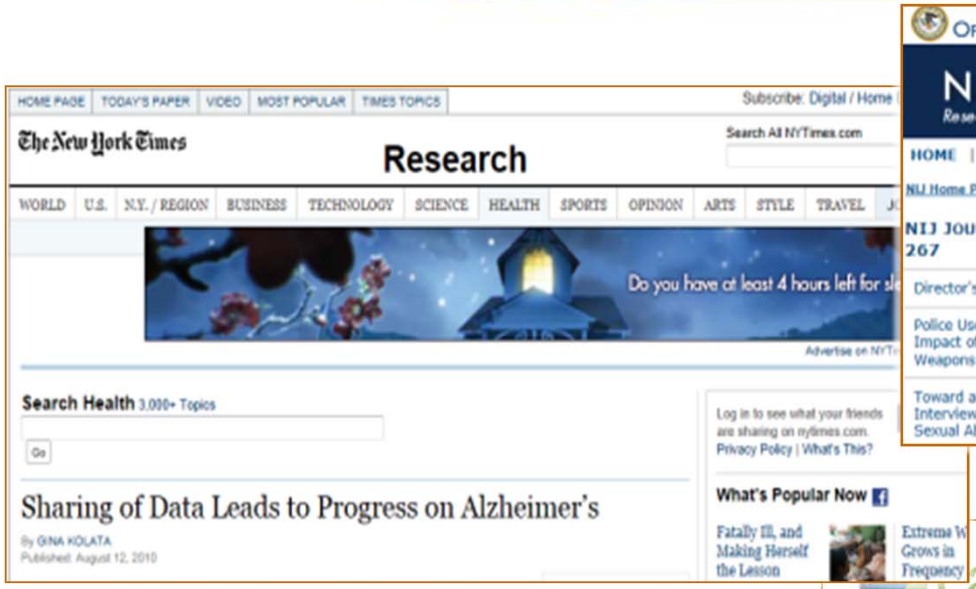
Dr. Francine Berman
Chair, RDA/US
Hamilton Distinguished Professor of Computer Science, RPI

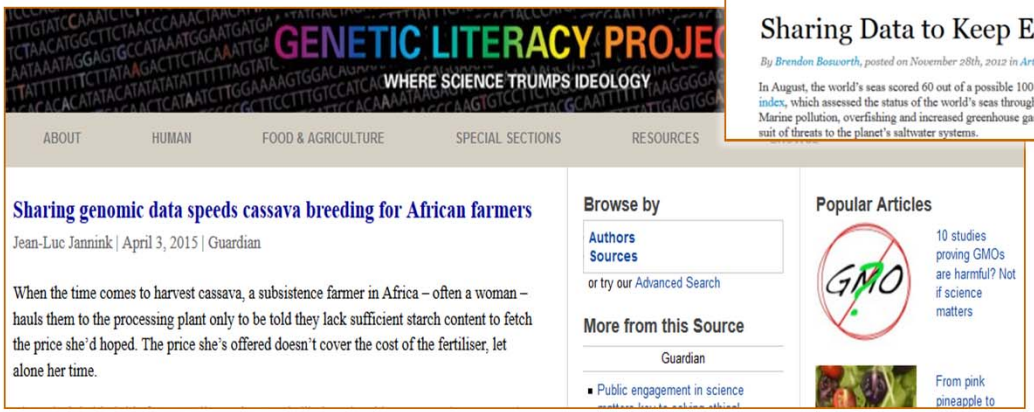
research data sharing without barriers
rd-alliance.org

The Research Data Alliance

- **Why do we need something like the RDA?**
- **What is the RDA?**
- **What does RDA produce?**
- **Who is the RDA community?**
- **What is RDA/US?**
- **On the horizon: International Data Week**

Why RDA? -- Data Sharing drives research and discovery in an information-rich world

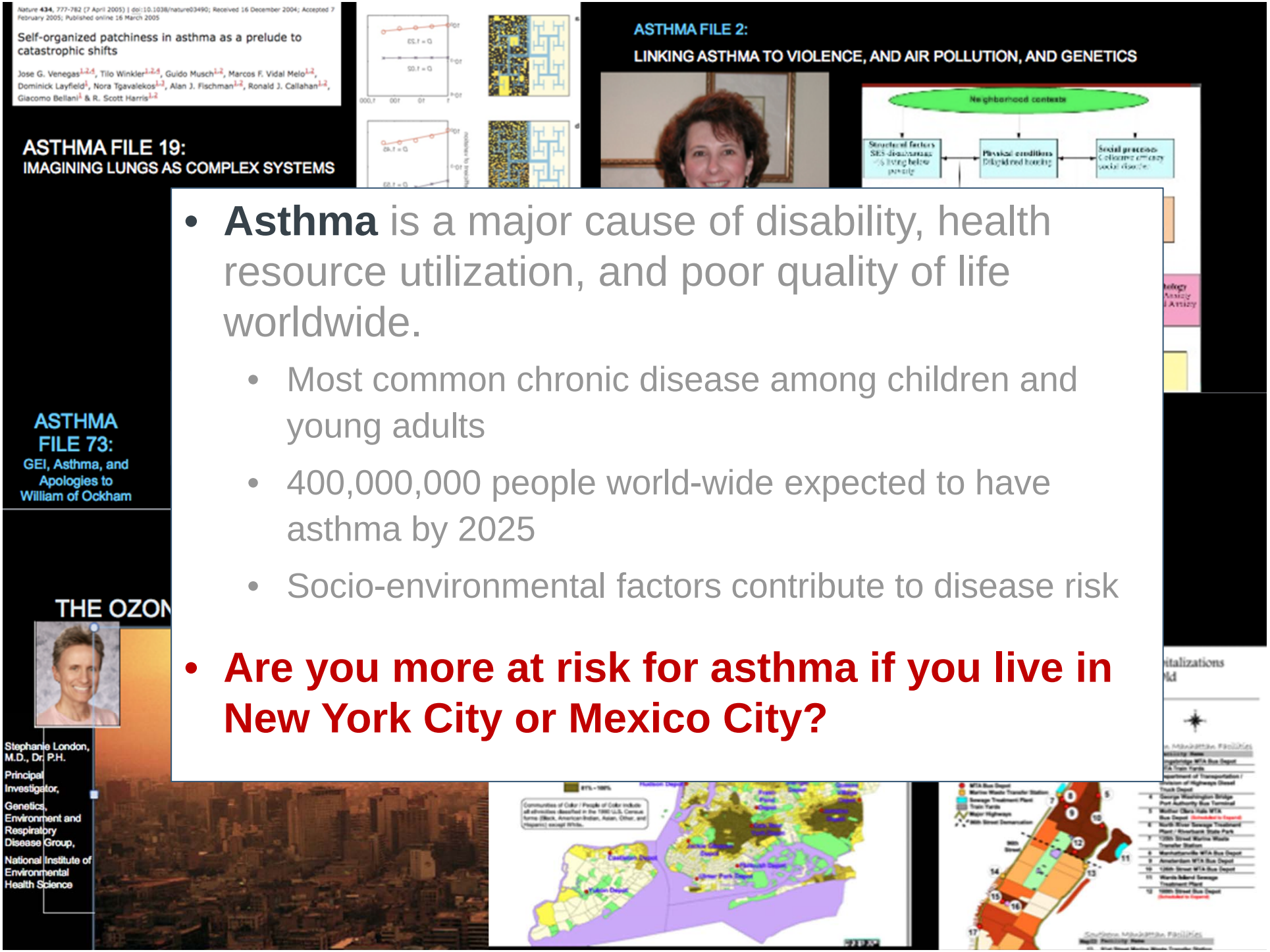




research data sharing without barriers
rd-alliance.org

Fran Berman

RDA
RESEARCH DATA ALLIANCE



- **Asthma** is a major cause of disability, health resource utilization, and poor quality of life worldwide.
  - Most common chronic disease among children and young adults
  - 400,000,000 people world-wide expected to have asthma by 2025
  - Socio-environmental factors contribute to disease risk
- **Are you more at risk for asthma if you live in New York City or Mexico City?**

Infrastructure needed to support data sharing

- **Making the data available isn't good enough.**
  - Data is not an asset if you don't know what it means.
  - Data is not useful if you can't find it.
  - Data needs to be in the right form for analysis.
  - Data needs to be preserved for results to be reproducible.

Both technical and social infrastructure needed to support data sharing

Adopted Policy

Systems Interoperability

Common Standards, Metadata

Sustainable Economics

Adopted Community Practice

Traffic Image: Mike Gonzalez

Training, Education, Workforce

research data sharing without barriers
rd-alliance.org

RESEARCH DATA ALLIANCE

Prioritizing infrastructure effort and investment challenging

The New York Times

Doctors Train to Spot Signs of A.D.H.D. in Children

Obama's Vote-Getting Tactics Struggle to Find the Uninsured

Obama Orders New Efficiency for Big Trucks

Public Defenders, Bolstered by a Work Analysis and Rulings, Push Back Against a Tide...

F.B.I. Joins Noose Is Le

U.S.

A Severe Winter Breaks Budgets as Well as Pipes

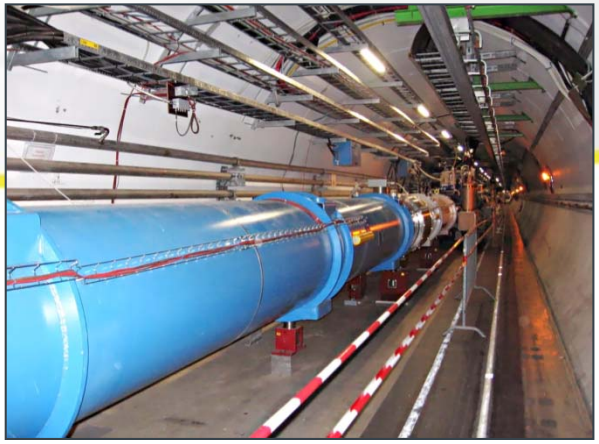
By JESSE McKINLEY and RICHARD PÉREZ-PEÑA FEB. 15, 2014

EMAIL
FACEBOOK
TWITTER
SAVE
MORE

SYRACUSE — Century-old water mains here have ruptured behind City Hall, popped in residential areas and split under the city's bar and restaurant district. The mayor says she has personally reported three breaks, while exhausted crews work 18-hour shifts in

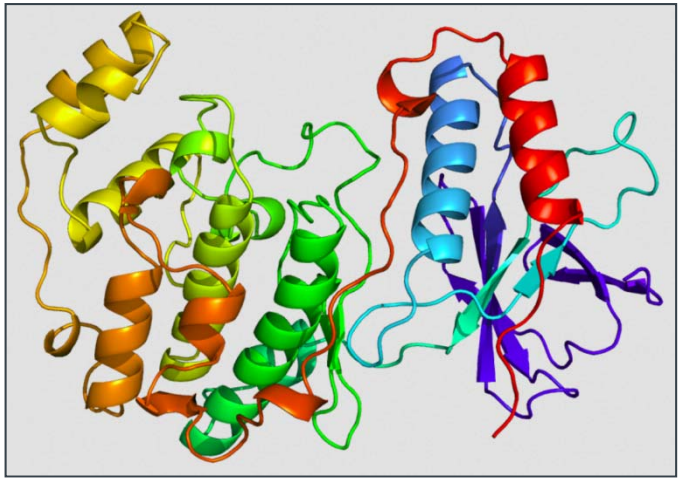
Stephanie A. Miner, the Syracuse mayor, said ***[infrastructure is] too often overlooked*** *when politicians want to spend money on economic development.*
"You don't cut ribbons for new water mains, but that's really what matters."
NY Times, Feburary 15, 2014

Organized infrastructure efforts can help



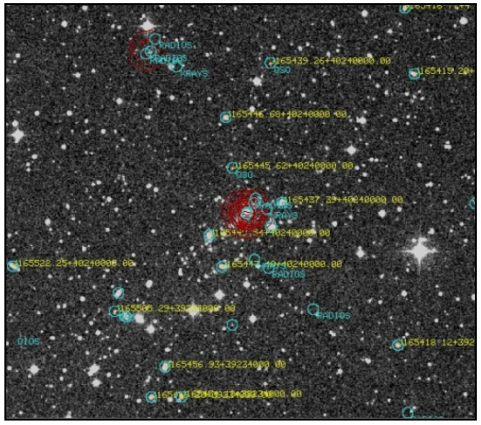
Large Hadron Collider; Wikimedia Commons, photo by Julian Herzog

High Energy Physics: **Data archival and analysis** part of LHC infrastructure plan and investment strategy



Protein Data Bank structure; Wikimedia Commons, public domain

Biomedical Science:
Agency and publisher's **policy** for deposit of structures in the PDB created an invaluable community resource.



Multi-wavelength data from various telescopes; Image courtesy of Robert Hanisch

Astronomy:
Development of **common practice and standards** support community analysis of astronomical databases and archives

The Research Data Alliance

- Why do we need something like the RDA?
- What is the RDA?
- What does RDA produce?
- Who is the RDA community?
- What is RDA/US?
- On the horizon: International Data Week

The Research Data Alliance

- **Research Data Alliance (RDA):** Global community-driven organization whose mission is to build the **social and technical** bridges (infrastructure) that enable data sharing.
- **Research Data Alliance Vision:** Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

RDA Approach: Solve Problems and Facilitate Progress

** Adoption of deliverables = incorporation, deployment, or implementation of infrastructure for use*

RDA Members come together as

- **Working Groups (WG)** – 12-18 month efforts to **build, adopt*, and use** specific pieces of infrastructure (deliverables)
- **Interest Groups (IG)** – longer-lived discussion forums that spawn Working Groups as specific pieces of needed infrastructure are identified.

RDA culture focuses on the pragmatic:

- **Working Groups must incorporate adopters** – no "build it and they will come"
- **Infrastructure must solve someone's problem** but not necessarily everyone's problems – not aiming for universal "esperanto" infrastructure
- **Maintain organizational agility** -- try things and improve them based on experience or drop them if they don't work
- **Promote technology-neutrality** -- RDA not a platform for specific infrastructure promotion or endorsement
- **Amplify impact** when possible
  - Proactively encourage **additional adopters**
  - **Collaborate with other organizations** to achieve their goals – RDA not looking for "world domination"

Deciding What Infrastructure to Build:
RDA Interest Groups as of June 2015

* in review

1. Agricultural Data
2. Active Data Management Plans*
3. Big Data
4. Biodiversity Data Integration
5. Brokering
6. Community Capability Model
7. Data Fabric
8. Data for Development
9. Data Foundations and Terminology*
10. Data in Context
11. Data Rescue
12. Development of cloud computing capacity and education in developing world research
13. Digital Practices in History and Ethnography
14. Domain Repositories Interest Group
15. Education and Training on handling of research data
16. ELIXIR Bridging Force
17. Engagement
18. Ethics and Social Aspects of Data
19. Federated Identity Management
20. Geospatial
21. Libraries for Research Data
22. Long tail of research data
23. Marine Data Harmonization
24. Metabolomics Data Interoperability
25. Metadata
26. National Data Services*
27. PID
28. Preservation e-Infrastructure
29. Quality of Urban Life
30. RDA/CODATA Legal Interoperability
31. RDA/CODATA Materials Data, Infrastructure & Interoperability
32. RDA/WDS Certification of Digital Repositories
33. RDA/WDS Publishing Data Cost Recovery for Data Centres
34. RDA/WDS Publishing Data
35. Repository Platforms for Research Data
36. Reproducibility
37. Research data needs of the Photon and Neutron Science community
38. Research Data Provenance
39. Service Management
40. Structural Biology
41. Vocabulary Services

Digital Practices in History and Ethnography IG

- **Focus:** Advance data standards, practices and infrastructure for historical and ethnographic research, digital humanities, and social sciences.
- **Platform shares** 2015-2016 with
  - Digital Himalaya Project (Cambridge, University of British Columbia)
  - Open Annotation Studio (MIT)
  - DARIAH (EU Digital Humanities)
- **"Issue Shares"** 2015-2016 with
  - Metadata Interest Group
  - Data Provenance Interest Group
  - Repository Platforms for Research Data Interest Group
  - Data Fabric Interest Group
- **Adopting data management infrastructure** from Practical Policy Working Group
- Developing **Working Group** proposal for metadata for the empirical humanities
- Community outreach: Helped organize the **RDA/US Digital Humanities Workshop**

Kim Fortun, U.S.

Mike Fortun, U.S.

Jason Baird Jackson, U.S.

RDA Working Group Deliverables – Beginning of a Pipeline (completion in fall 2014)

Working Group	Deliverable	Impact	Adopters
Data Foundation and Terminology Working Group	Basic vocabulary of foundational terminology, query tool	Ensures researchers use a common terminology when referring to data	EUDAT, DKRZ, Deep Carbon Observatory, CLARIN, EPOS
Data Type Registries Working Group	Data type model and prototype registry	Provides machine-readable and researcher-accessible registries of data types that support the accurate use of data	CNRI, International DOI Foundation, Materials Genome Initiative, Deep Carbon Observatory
PID Information Types Working Group	Persistent identifier registry	Conceptual model for structuring typed information to better identify PIDs, common interface for access to this information	Materials Genome Initiative, Deep Carbon Observatory, Data Conservancy, DKRZ
Practical Policy Working Group	Basic set of machine actionable rules	Policy templates that can be used to support data sharing and interchange between communities	Platform for Experimental Collaborative Ethnography, EUDAT, Washington University St. Louis, RENCI, DataNet Federation Consortium, CESNET, Odum Inst.

RDA Working Group Deliverables – Beginning of a Pipeline (completion in spring 2015)

Working Group	Deliverables	Impact	Adopters
Data Citation Working Group	Dynamic-data citation methodology that supports efficient processing of data and linking from publications	Researchers can reference precise subsets of changing data	NERC, ESIP, CLARIN, Virtual Atomic and Molecular Data Centre
Metadata Standards Directory Working Group	Prototype Metadata Standards Directory and use cases	Information can be maintained transparently and with full version control.	Digital Curation Centre, JISC, DataOne
Wheat Data Interoperability Working Group	Common framework for Wheat Data Terminology to enable interoperability between distinct data collections	Semantically linked terms describing wheat data so researchers can share harvest and related information between data sets and communities	Wheat Initiative Information System, FAO AIMS, INRA
Data Description Registry Interoperability Working Group	Systems and graph technologies to link data across multiple registries to facilitate search and discovery	Enables more efficient discovery of data sets	Australian National Data Service, CERN, DANS, DataCite, DataPASS, Thomson Reuters, Cornell

Data Citation Working Group


Andreas Rauber, Austria


Ari Asmi, Finland

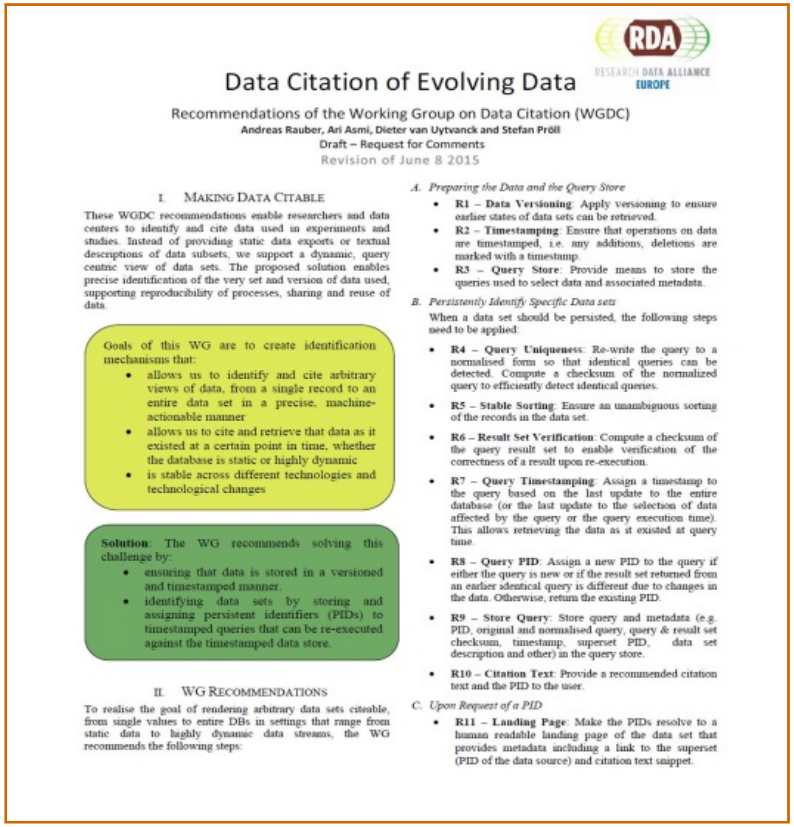

Dieter van Uytvanck, Netherlands

Problem: Research data is dynamic. Data sets change when

- New data is added
- Errors are corrected
- Data is re-ordered, etc.

How can you repeat an experiment based on a dataset that keeps growing and changing?

How do you identify and cite precisely the subset of dynamic data used in a study?



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Draft – Request for Comments

Revision of June 8 2015

I. MAKING DATA CITABLE

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very set and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

II. WG RECOMMENDATIONS

To realise the goal of rendering arbitrary data sets citeable, from single values to entire DBs in settings that range from static data to highly dynamic data streams, the WG recommends the following steps:

A. Preparing the Data and the Query Store

- **R1 – Data Versioning**: Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping**: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store**: Provide means to store the queries used to select data and associated metadata.

B. Persistently Identify Specific Data sets

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness**: Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.
- **R5 – Stable Sorting**: Ensure an unambiguous sorting of the records in the data set.
- **R6 – Result Set Verification**: Compute a checksum of the query result set to enable verification of the correctness of a result upon re-execution.
- **R7 – Query Timestamping**: Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at query time.
- **R8 – Query PID**: Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID.
- **R9 – Store Query**: Store query and metadata (e.g. PID, original and normalised query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store.
- **R10 – Citation Text**: Provide a recommended citation text and the PID to the user.

C. Upon Request of a PID

- **R11 – Landing Page**: Make the PIDs resolve to a human readable landing page of the data set that provides metadata including a link to the superset (PID of the data source) and citation text snippet.

Data Citation Working Group

- **What the WG is doing:**
  - Focusing on the problems of **identifying and citing data within large, dynamic (changing) datasets** in a machine-actionable manner
  - Solution approach focuses on data versioning, data timestamping and data identification. Approach is DBMS- and technology- independent
- **Deliverables:** 13 Recommendations dealing with
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Response to request of a PID
  - Modifications to / migration of the data infrastructure

- **What the WG is *not* doing:**
  - Developing PID systems, developing specific metadata categories, new approaches to attribution.
  - Starting from scratch: WG leveraging work from other RDA WG and community efforts on data citation

RDA
RESEARCH DATA ALLIANCE

Data Citation WG Adopters

- **Pilot workshops and implementations by**
  - Various EU projects (TIMBUS, SCAPE,…)
  - NERC (UK Natural Environment Research Council Data Centres)
  - ESIP (Earth Science Information Partners)
  - CLARIN (XML, Field Linguistics Transcriptions)
  - Virtual Atomic and Molecular Data Centre
- **Prototype solutions for**
  - SQL, CSV, XML
  - LOD/RDF, triple-store DBs in the queue
  - Distributed data
  - Video of CSV prototype available at http://datacitation.eu

Pragmatic Progress:

- Group published approach 2013 and used **RDA as a vehicle** to develop real infrastructure based on peer-reviewed, vetted ideas
- **Broader set collaborators, adopters, domains** helped transition effort into **needed community infrastructure**

RDA Wheat Data Interoperability WG

Esther Dzalé Yeumo, France

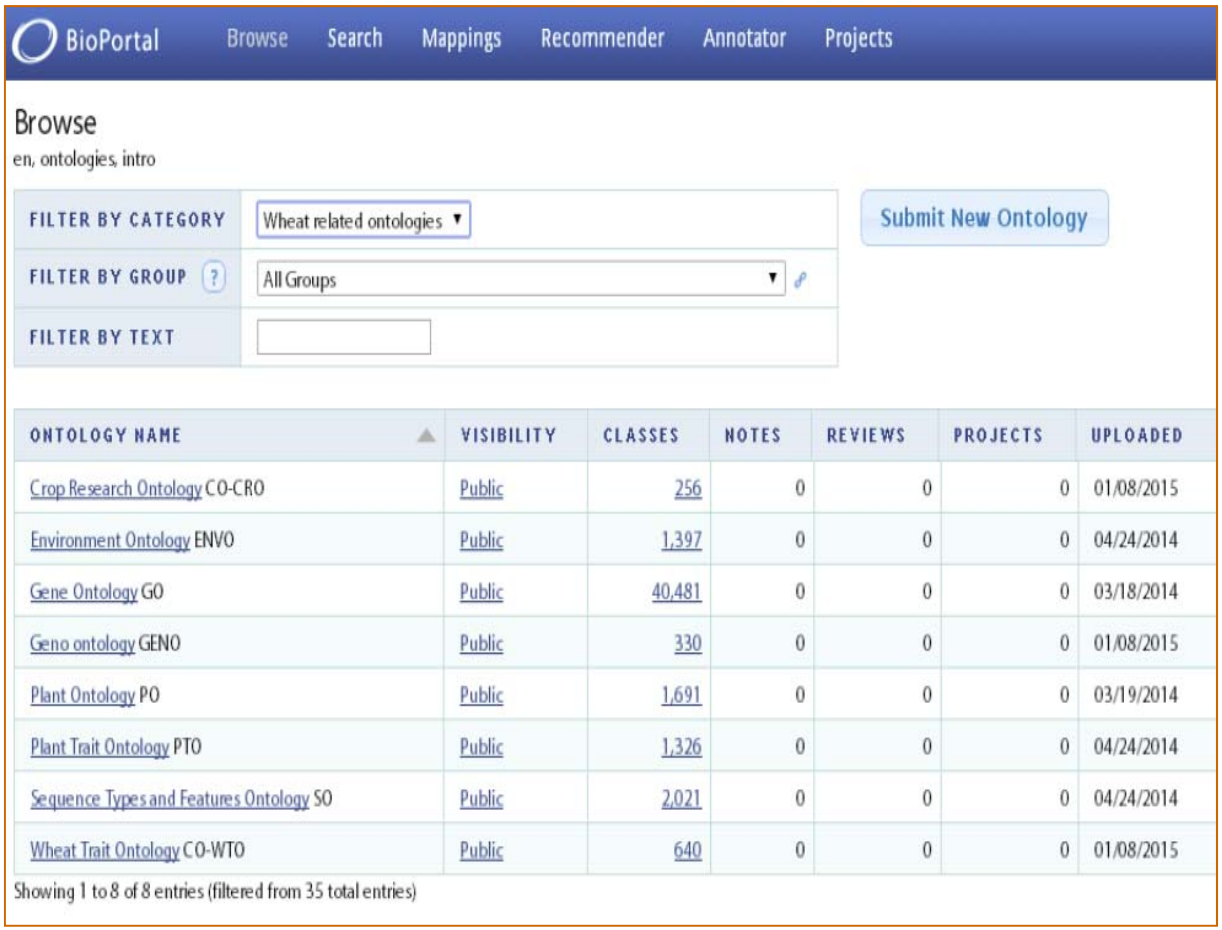
Richard Fulss, Mexico

- **Focus:** Agricultural productivity to feed the planet is a major societal challenge. What data interoperability can be developed to help address agricultural productivity challenges?
- **Solution approach:** Make critical data sets for agricultural interoperable by agreeing on a common set of
  - Metadata standards
  - Data formats
  - Vocabularies
  - Guidelines for distributing, representing, and linking data

What the WG is doing:

- WG building an interactive "cookbook" with recommendations and guidelines on data format and standards
- Developing common wheat-related vocabularies and including them in a human and machine-readable bio-portal
- Building a prototype interoperability framework for specific use cases.

WG enabling more effective agricultural research



BioPortal Browse Search Mappings Recommender Annotator Projects

Browse

en, ontologies, intro

FILTER BY CATEGORY: Wheat related ontologies

FILTER BY GROUP: All Groups

FILTER BY TEXT

Submit New Ontology

ONTOLOGY NAME	VISIBILITY	CLASSES	NOTES	REVIEWS	PROJECTS	UPLOADED
Crop Research Ontology CO-CRO	Public	256	0	0	0	01/08/2015
Environment Ontology ENVO	Public	1,397	0	0	0	04/24/2014
Gene Ontology GO	Public	40,481	0	0	0	03/18/2014
Geno ontology GENO	Public	330	0	0	0	01/08/2015
Plant Ontology PO	Public	1,691	0	0	0	03/19/2014
Plant Trait Ontology PTO	Public	1,326	0	0	0	04/24/2014
Sequence Types and Features Ontology SO	Public	2,021	0	0	0	04/24/2014
Wheat Trait Ontology CO-WTO	Public	640	0	0	0	01/08/2015

Showing 1 to 8 of 8 entries (filtered from 35 total entries)

Adoption and next steps:

- Framework will be incorporated into the **Wheat Information System** of the Global Wheat Initiative, **Coherence in Information for Agricultural Research for Development** (CIARD), etc.
- *Subsequent work:* Framework will be adapted to other crops such as **Rice** and **Maize.**

Portal image from the RDA Outputs booklet, 2015 https://rd-alliance.org/rda-outputs.html

Fran Berman

RDA RESEARCH DATA ALLIANCE

RDA Plenaries – **Participatory** Global Community Gatherings and Working Meetings

March 2013

240 participants from 30 countries

first Working Groups and Interest Groups

September 2014

550 participants from 40 countries

first RDA Deliverables

September 2013

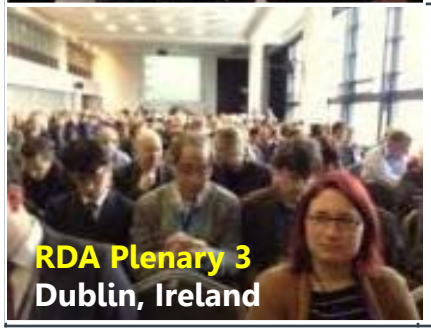
380 participants from 32 countries

first Birds-of-a-Feather sessions

March 2015

240 participants from 30 countries,

first Adoption Day



March 2014

497 participants from 22 countries

first Organizational Assembly

RDA
RESEARCH DATA ALLIANCE

RDA Plenary 6
Paris, France

September 2015

RDA Plenary 6

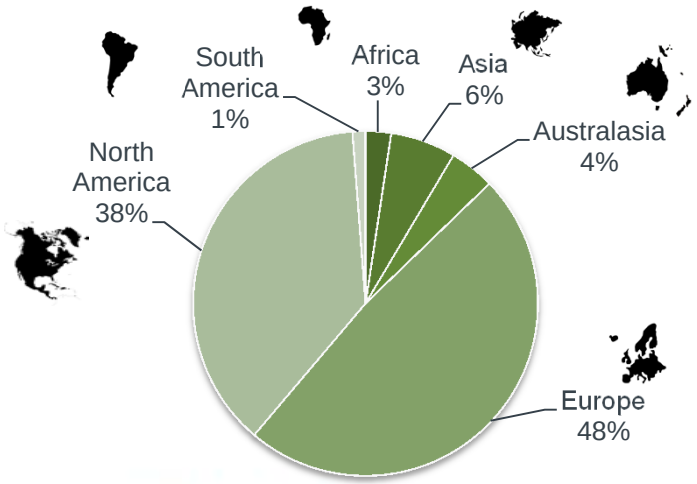
first RDA data challenge (climate data)

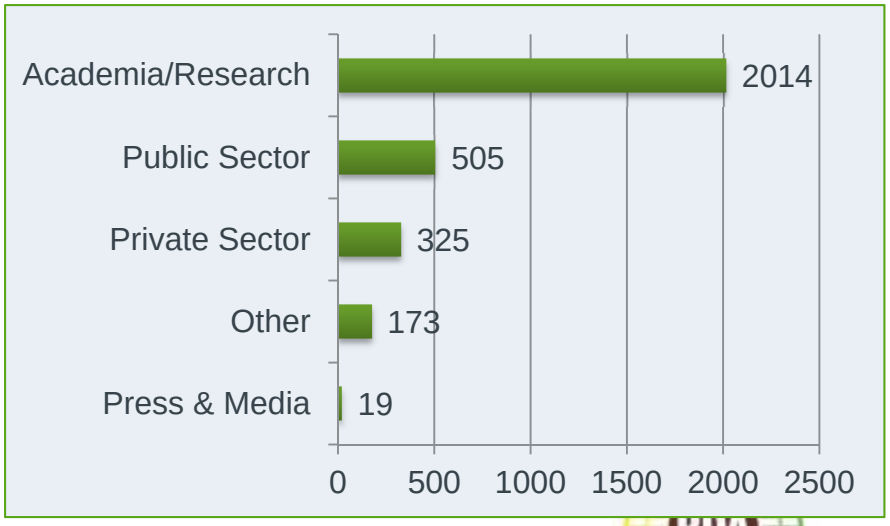
Co-location with RDA Plenaries helping build linkages within the data community

Who is RDA?

Total RDA Community as of July, 2015:

3029 RDA members from 103 countries



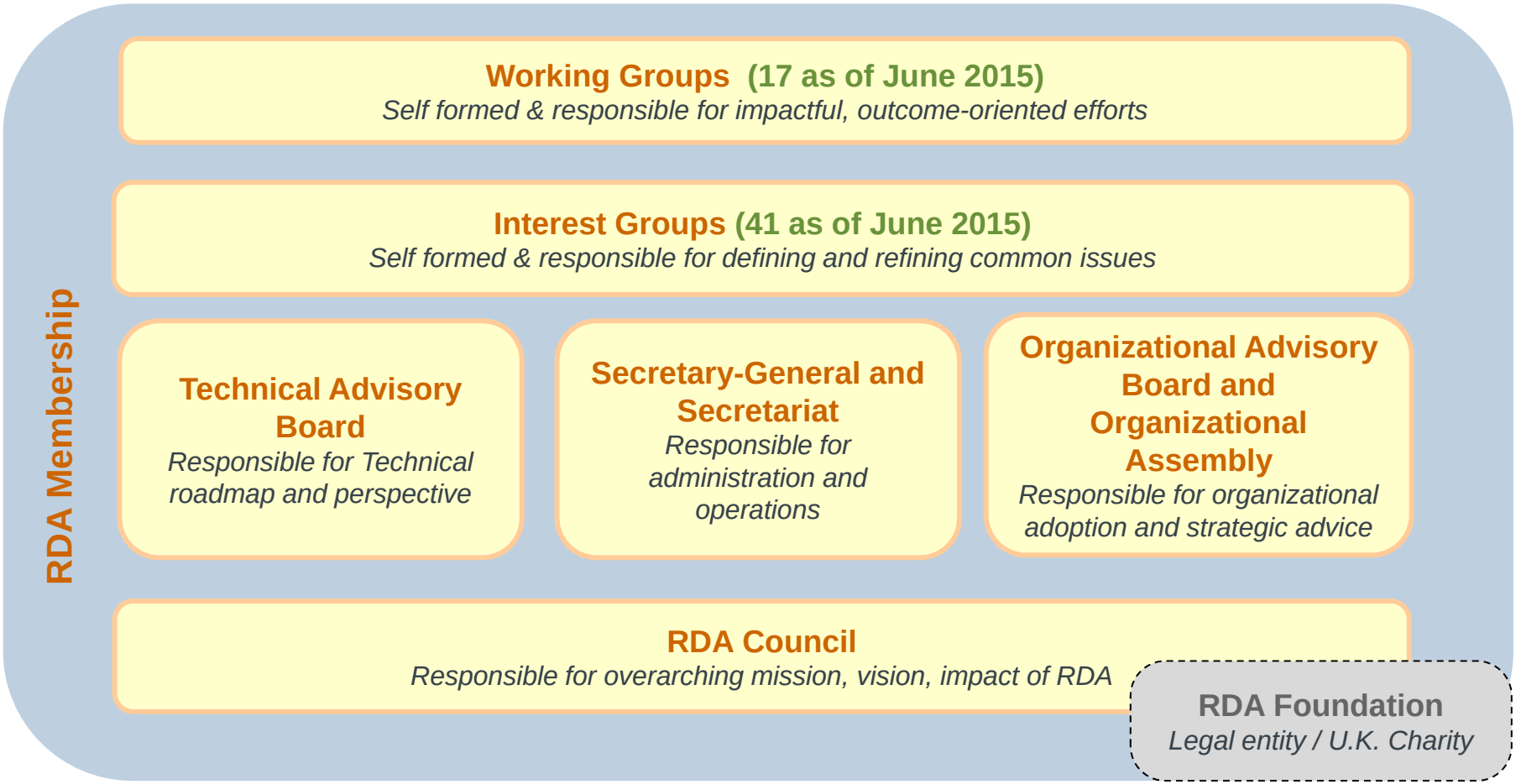


research data sharing without barriers
rd-alliance.org

Fran Berman

RDA
RESEARCH DATA ALLIANCE

How RDA is Organized



How are we doing? -- RDA early evaluation

- **Really working well:**
  - RDA delivering on its mission: groups really building infrastructure that people need, adopters incorporating infrastructure
  - RDA Plenaries proving to be outstanding drivers for building broad, diverse, synergistic data community
    - RDA becoming a “go-to” venue for people to come and solve real problems
    - Peripheral meetings adding / getting tremendous value
- **More complicated than we’d planned for**
  - How do we maximize the impact of deliverables beyond initial adopters? Who maintains them? Who evolves them?
  - How do we support community growth and activity with a small, lightweight governance structure?
  - What are the right partnership models with startup / medium / large organizations?
- **Really challenging**
  - How do we sustain the organization beyond the current set of funding streams?

Emerging **RDA Value Proposition**

- **RDA value proposition for individuals and researchers:**
  - Expanded, global collaboration network with diverse perspectives that can help vet/improve work
  - Vehicle for accelerating the development of infrastructure needed to drive discovery
  - Engagement with a broader set of domains and stakeholders
- **RDA value proposition for communities, organizations, private sector**
  - Opportunities for engagement with an expanded and diverse global network of potential collaborators, partners, employees
  - Vehicle to accelerate the incorporation of data sharing technologies
- **RDA value proposition for countries / public sector**
  - Vehicle for promoting leadership and competitiveness of national research communities within a global environment
  - Vehicle for accelerating the development of national and global infrastructure needed to accelerate discovery
  - Vehicle for strengthening international, inter-disciplinary and inter-sector collaborations

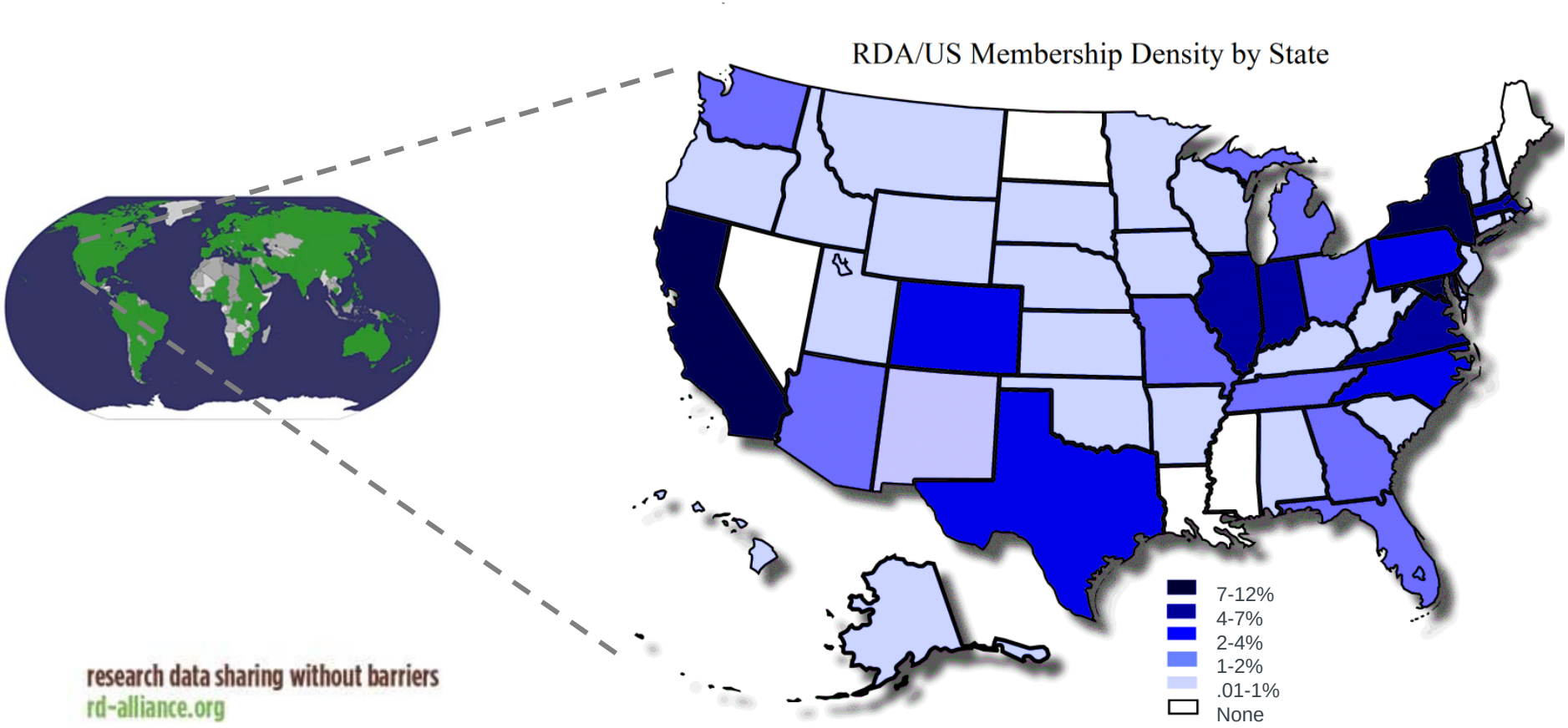
The Research Data Alliance

- Why do we need something like the RDA?
- What is the RDA?
- What does RDA produce?
- Who is the RDA community?
- What is RDA/US?
- On the horizon: International Data Week

RDA/US = U.S. members of RDA

RESEARCH DATA ALLIANCE
UNITED STATES

- **RDA/US Mission:** To build RDA community in the U.S. and leverage RDA momentum to advance the U.S. data community
- **Currently ~1000 members of RDA in 45 states**



Selected RDA/US Activities for 2015-2016

- **Student / Early Career** program
- Targeted **Outreach Workshops** with data-enabled communities and organizations
- Development of an RDA/US **Ambassador Program** to strengthen ties with communities and disciplines
- Continuation of **Joint Partnership Agreements** between RDA/US and U.S.-based organizations to co-sponsor activities and events that build the RDA community
- Planning for **Plenary 8** and International Data Week
- **Adoption Amplification** seed projects for RDA deliverables
- **Testbed:** Proposed creation of a testbed for RDA deliverables on use cases from the RDA community
- Hosting and U.S. participant support for **WG Coordination meetings** and "data fabric" development
- Development of **RDA/US website** and **communications**, publications, curriculum, collateral for RDA/US and data sharing

RDA/US Student / Early Career Programs

- **RDA/US Student / Early Career Programs**
  - Expand/strengthen the professional network of Fellows and Interns
  - Build/strengthen the generational pipeline within the data community
  - Build linkages within communities
- **NSF RDA/US Fellows and Interns Pilot 2013-2015:**
  - Work with specific Interest or Working Groups, report on work at Plenaries
- **Sloan Foundation DataShare Program 2015-2018**
  - 12-18 month student/early career projects focus on evaluating, trial use, or improvement of products developed within a Working Group, developing and testing adoption strategies, or facilitation of interaction between RDA groups.

Current funding for RDA/US activities and organization

Data SHARE
Sloan Foundation
(U.S. Student and early career engagement and support)

COORDINATION / OUTREACH
NIST
(meeting and U.S. participant support for WG coordination meetings)

RDA2
NSF
(RDA/US organizational support, RDA/US in-region contributions to RDA Secretariat, RDA/US community building, RDA/US pilots)

RDA1
NSF
(RDA/US organizational support, RDA/US in-region contributions to RDA Secretariat, RDA/US community building)

FY 2013 FY2014 FY2015 FY2016 FY2017 FY2018 FY2019

On the Horizon: Upcoming RDA Plenaries and International Data Week

- **Plenary 6:** September 22-25 2015, Paris France
- **Plenary 7:** March 1-3, 2016, Tokyo Japan
- **Plenary 8:** September 11-16. 2016, Washington DC
  - RDA/US collaborating with CODATA and International Council for Science World Data System to co-host "International Data Week" in September, 2016 in Washington, D.C.

- **Everyone is welcome to RDA Plenaries**
- Let us know if your communities will want time / space for Birds-of-a-Feather meetings, Workshops, etc.

Getting involved with RDA

Ways to Engage	How	Benefits	Contribution
Join RDA as an individual member	Sign up at rd-alliance.org. Agree to RDA principles.	Diverse global network of collaborators / colleagues. Vehicle for building and adopting data sharing infrastructure	Participate in Working and Interest Groups, discussion and evaluation ($0)
Join RDA as an Organizational Member or an Organizational Affiliate	Member: Contact Juan Bicarregui or Walter Stewart at rd-alliance.org Affiliate: Contact Secretary General Mark Parsons. Requires approval of Council	Synergistic collaboration with RDA advances organization's mission, and provides opportunities for early adoption of RDA deliverables and building of data infrastructure	Member: Evaluate and advise on RDA deliverables and organizational approach; ($1K-$10K/ year dues depending on organizational size.) Affiliate: Co-sponsor and collaborate on activities ($0).
Join RDA's Funders Forum	Contact Fran Berman and John Wood (Council co-Chairs)	Synergistic coordination with international agencies and non-profits to leverage / coordinate global efforts and support RDA organization and community	Contribute to community and/or organization. ($Variable. See Fran for details.)
Engage with RDA/US to support U.S. RDA community	Contact Fran Berman (RDA/US Chair) or Kathy Fontaine (RDA/US Managing Director)	Convene community and build needed infrastructure through workshops, adoption efforts, etc.	Contribute "in region" to U.S. RDA community and/or organization. ($Variable. See Fran for details.)

Thank you

bermaf@rpi.edu

research data sharing without barriers

rd-alliance.org

Working Groups focusing on both Technical and Social Infrastructure