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
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?
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

- **NO SMOKING**
  - Adopted Policy

- **Common Standards, Metadata**
  - Systems Interoperability

- **Sustainable Economics**

- **Traffic Image**
  - Adopted Community Practice
    - Traffic Image: Mike Gonzalez

- **Training, Education, Workforce**
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, Februrary 15, 2014
Organized infrastructure efforts can help

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

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

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

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”

*Adoption of deliverables = incorporation, deployment, or implementation of infrastructure for use*
Deciding What Infrastructure to Build: RDA Interest Groups as of June 2015

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

* in review
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)

<table>
<thead>
<tr>
<th>Working Group</th>
<th>Deliverable</th>
<th>Impact</th>
<th>Adopters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Foundation and Terminology Working Group</td>
<td>Basic vocabulary of foundational terminology, query tool</td>
<td>Ensures researchers use a common terminology when referring to data</td>
<td>EUDAT, DKRZ, Deep Carbon Observatory, CLARIN, EPOS</td>
</tr>
<tr>
<td>Data Type Registries Working Group</td>
<td>Data type model and prototype registry</td>
<td>Provides machine-readable and researcher-accessible registries of data types that support the accurate use of data</td>
<td>CNRI, International DOI Foundation, Materials Genome Initiative, Deep Carbon Observatory</td>
</tr>
<tr>
<td>PID Information Types Working Group</td>
<td>Persistent identifier registry</td>
<td>Conceptual model for structuring typed information to better identify PIDs, common interface for access to this information</td>
<td>Materials Genome Initiative, Deep Carbon Observatory, Data Conservancy, DKRZ</td>
</tr>
<tr>
<td>Practical Policy Working Group</td>
<td>Basic set of machine actionable rules</td>
<td>Policy templates that can be used to support data sharing and interchange between communities</td>
<td>Platform for Experimental Collaborative Ethnography, EUDAT, Washington University St. Louis, RENCI, DataNet Federation Consortium, CESNET, Odum Inst.</td>
</tr>
<tr>
<td>Working Group</td>
<td>Deliverables</td>
<td>Impact</td>
<td>Adopters</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Data Citation Working Group</td>
<td>Dynamic-data citation methodology that supports efficient processing of data and linking from publications</td>
<td>Researchers can reference precise subsets of changing data</td>
<td>NERC, ESIP, CLARIN, Virtual Atomic and Molecular Data Centre</td>
</tr>
<tr>
<td>Metadata Standards Directory Working Group</td>
<td>Prototype Metadata Standards Directory and use cases</td>
<td>Information can be maintained transparently and with full version control.</td>
<td>Digital Curation Centre, JISC, DataOne</td>
</tr>
<tr>
<td>Wheat Data Interoperability Working Group</td>
<td>Common framework for Wheat Data Terminology to enable interoperability between distinct data collections</td>
<td>Semantically linked terms describing wheat data so researchers can share harvest and related information between data sets and communities</td>
<td>Wheat Initiative Information System, FAO AIMS, INRA</td>
</tr>
<tr>
<td>Data Description Registry Interoperability Working Group</td>
<td>Systems and graph technologies to link data across multiple registries to facilitate search and discovery</td>
<td>Enables more efficient discovery of data sets</td>
<td>Australian National Data Service, CERN, DANS, DataCite, DataPASS, Thomson Reuters, Cornell</td>
</tr>
</tbody>
</table>
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?
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
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**
**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

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 Plenaries – Participatory Global Community Gatherings and Working Meetings

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Participants</th>
<th>Location</th>
<th>Deliverables/Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2013</td>
<td>240 participants from 30 countries, first Working Groups and Interest Groups</td>
<td>240</td>
<td>Gothenburg, Sweden</td>
<td></td>
</tr>
<tr>
<td>September 2013</td>
<td>380 participants from 32 countries, first Birds-of-a-Feather sessions</td>
<td>380</td>
<td>Washington, DC</td>
<td></td>
</tr>
<tr>
<td>March 2014</td>
<td>497 participants from 22 countries, first Organizational Assembly</td>
<td>497</td>
<td>Dublin, Ireland</td>
<td></td>
</tr>
<tr>
<td>September 2014</td>
<td>550 participants from 40 countries, first RDA Deliverables</td>
<td>550</td>
<td>San Diego, California</td>
<td></td>
</tr>
<tr>
<td>March 2015</td>
<td>240 participants from 30 countries, first Adoption Day</td>
<td>240</td>
<td>Paris, France</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Month</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>May - July 13</td>
<td>393</td>
</tr>
<tr>
<td>Aug - Oct 13</td>
<td>993</td>
</tr>
<tr>
<td>Nov - Jan 14</td>
<td>1276</td>
</tr>
<tr>
<td>Feb - Apr 14</td>
<td>1658</td>
</tr>
<tr>
<td>May - July 14</td>
<td>2051</td>
</tr>
<tr>
<td>Aug - Oct 15</td>
<td>2407</td>
</tr>
<tr>
<td>Nov - Jan 15</td>
<td>2639</td>
</tr>
<tr>
<td>Feb - Apr 15</td>
<td>2851</td>
</tr>
<tr>
<td>May - Jun 15</td>
<td>2936</td>
</tr>
</tbody>
</table>

- Europe: 48%
- North America: 38%
- Australasia: 4%
- Asia: 6%
- Africa: 3%
- South America: 1%

Research data sharing without barriers
rd-alliance.org

Press & Media

<table>
<thead>
<tr>
<th>Sector</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academia/Research</td>
<td>2014</td>
</tr>
<tr>
<td>Public Sector</td>
<td>505</td>
</tr>
<tr>
<td>Private Sector</td>
<td>325</td>
</tr>
<tr>
<td>Other</td>
<td>173</td>
</tr>
<tr>
<td>Press &amp; Media</td>
<td>19</td>
</tr>
</tbody>
</table>

Fran Berman
How RDA is Organized

RDA Membership

- **Working Groups (17 as of June 2015)**
  Self formed & responsible for impactful, outcome-oriented efforts

- **Interest Groups (41 as of June 2015)**
  Self formed & responsible for defining and refining common issues

- **Technical Advisory Board**
  Responsible for Technical roadmap and perspective

- **Secretary-General and Secretariat**
  Responsible for administration and operations

- **Organizational Advisory Board and Organizational Assembly**
  Responsible for organizational adoption and strategic advice

- **RDA Council**
  Responsible for overarching mission, vision, impact of RDA

RDA Funders Forum
Support for RDA Organization and Community
(Currently NSF, NIST, Sloan Foundation, European Commission, Australian Government)

RDA Foundation
Legal entity / U.K. Charity
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

- **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

- 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)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fran Berman
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

<table>
<thead>
<tr>
<th>Ways to Engage</th>
<th>How</th>
<th>Benefits</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Join RDA as an individual member</td>
<td>Sign up at rd-alliance.org. Agree to RDA principles.</td>
<td>Diverse global network of collaborators / colleagues. Vehicle for building and adopting data sharing infrastructure</td>
<td>Participate in Working and Interest Groups, discussion and evaluation ($0)</td>
</tr>
</tbody>
</table>
| 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.) |

Fran Berman
Thank you
bermaf@rpi.edu

research data sharing without barriers
rd-alliance.org
Working Groups focusing on both Technical and Social Infrastructure

Social infrastructure

- Repository Audit and Certification DSA–WDS Partnership
- Brokering Governance
- PID Information Types
- Data Type Registries

Technical infrastructure

- Data Foundation and Terminology
- Practical Policy
- RDA/WDS Publishing Data Workflows
- RDA/WDS Publishing Data Bibliometrics
- RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
- Metadata Standards Directory
- Data Description Registry
- Interoperability
- BioSharing Registry
- Wheat Data Interoperability

Data providers

Beneficiary dimension

Data consumers

Social infrastructure

Technical infrastructure

Data providers

Beneficiary dimension

Data consumers

Repository Audit and Certification DSA–WDS Partnership

Brokering Governance

PID Information Types

Data Type Registries

Data Foundation and Terminology

Practical Policy

RDA/WDS Publishing Data Workflows

RDA/WDS Publishing Data Bibliometrics

RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World

Metadata Standards Directory

Data Description Registry

Interoperability

BioSharing Registry

Wheat Data Interoperability

RDA/WDS Publishing Data Services

Research data sharing without barriers

rd-alliance.org

Fran Berman

RDA
RESEARCH DATA ALLIANCE