Open Data is Not Enough
Making Data Sharing Work

Mark A. Parsons
0000-0002-7723-0950
Secretary General

RDA meets the Italian scientific community
Pisa, Italy
14 July 2015
September Arctic Sea Ice Extent, 1979 to 2012

September 2012  September 1979

Credit: National Snow and Ice Data Center, NASA Earth Observatory

NSIDC
Arctic Sea Ice Extent
(Area of Ocean with at least 15% sea ice)
The many representations (and conceptions) of sea ice

Evolution of sea ice data products at NSIDC, presented by Ruth Duerr
March 10, 2015, RDA 5th Plenary, San Diego
Figure courtesy Donna Scott
The generative value of data

• Generative value per Jonathan Zittrain (2008) as interpreted and extended to data by John Wilbanks:

  “the capacity to produce unanticipated change through unfiltered contributions from broad and varied audiences.” —J. Zittrain

• Data become more generative by being more adaptable, more easily mastered, more accessible, and more connected and influential.

• Not net present value but net potential value.
To make open work we need

- **Curation**—increasing the (generative) value of the data
- **Context**—of both data and application of provider and user
- **Trust**—of data, information, organisations, institutions,….
- Interfaces, connections, relationships, mediation—**Bridges**
- **People**
Research Data Alliance

Vision
Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

Mission
RDA builds the social and technical bridges that enable open sharing of data.
Dynamics of Infrastructure

- Infrastructures become “ubiquitous, accessible, reliable, and transparent” as they mature.

- Systems → Networks → Inter-networks

  - “system-building, characterized by the deliberate and successful design of technology-based services.”

  - “technology transfer across domains and locations results in variations on the original design, as well as the emergence of competing systems.”

  - Finally, “a process of consolidation characterized by gateways that allow dissimilar systems to be linked into networks.”
Not what, but

When is infrastructure?
Not what, but

When and

Who is infrastructure?
Bridges and Gateways

Gateways are often wrongly understood as “technologies,” i.e. hardware or software alone. A more accurate approach conceives them as combining a technical solution with a social choice, i.e. a standard, both of which must be integrated into existing users’ communities of practice. Because of this, gateways rarely perform perfectly.

— Edwards et al. 2007
Infrastructure is

Relationships, interactions, and connections between people, technologies, and institutions
“Create - Adopt - Use”
(in 12-18 months)

- Adopted Policy
- Systems Interoperability
- Common Types, Standards, Metadata
- Sustainable Economics
- Adopted Community Practice
- Training, Education, Workforce

Traffic Image: Mike Gonzalez

Fran Berman, Research Data Alliance
Shared Principles

• Openness
• Consensus
• Balance
• Harmonization
• Community Driven
• Non-profit
RDA: Accelerate Data Sharing and Interoperability Across Cultures, Communities, Scales, Technologies

- **Technical parts of the data engine:**
  - Data type registries reference model
  - Wheat data interoperability framework

- **Rules of the road:**
  - Common agreement on data citation
  - Common practice for data repositories
  - Principles of legal interoperability

- **Better drivers**
  - Summer schools in data science and cloud computing in the developing world (with CODATA)
  - Active data management plan development and monitoring
Initial Products—adopt one today!

• A basic vocabulary of foundational terminology and query tool to make sure we know what we’re talking about.

• A data type model and registry (‘‘MIME-types’’ for data) to help tools interpret, display, and process data.

• A persistent identifier type registry to help search engines understand what they are pointing to and retrieving.

• A basic set of machine actionable rules to enhance trust

• Coming soon:
  • A metadata standards directory so we can describe similar things consistently
  • A dynamic-data citation methodology so we can reference precise subsets of changing data.
  • Semantically linked terms describing wheat data so we can share harvest and related information around the world
  • Services and methods for finding data across multiple registries, to help cross disciplinary and multi-facetted discovery.
  • A unified repository certification scheme to reduce confusion and improve trust.
RDA Working Groups

1. Brokering Governance
2. Data Citation WG
3. Data Description Registry Interoperability
4. Data Foundation and Terminology WG
5. Data Type Registries WG
7. PID Information Types WG
8. Practical Policy WG
9. RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
10. RDA/WDS Publishing Data Bibliometrics WG
11. RDA/WDS Publishing Data Services WG
12. RDA/WDS Publishing Data Workflows WG
13. Repository Audit and Certification DSA–WDS Partnership WG
14. Repository Platforms for Research Data*
15. The BioSharing Registry: connecting data policies, standards & databases in life sciences*
16. Wheat Data Interoperability WG

* in review
# RDA Interest Groups

<table>
<thead>
<tr>
<th>No.</th>
<th>Interest Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Agricultural Data Interoperability IG</td>
</tr>
<tr>
<td>2.</td>
<td>Active Data Management Plans*</td>
</tr>
<tr>
<td>3.</td>
<td>Big Data Analytics IG</td>
</tr>
<tr>
<td>4.</td>
<td>Biodiversity Data Integration IG</td>
</tr>
<tr>
<td>5.</td>
<td>Brokering IG</td>
</tr>
<tr>
<td>6.</td>
<td>Community Capability Model IG</td>
</tr>
<tr>
<td>7.</td>
<td>Data Fabric IG</td>
</tr>
<tr>
<td>8.</td>
<td>Data for Development</td>
</tr>
<tr>
<td>9.</td>
<td>Data Foundations and Terminology IG*</td>
</tr>
<tr>
<td>10.</td>
<td>Data in Context IG</td>
</tr>
<tr>
<td>11.</td>
<td>Development of cloud computing capacity and education in developing world research</td>
</tr>
<tr>
<td>12.</td>
<td>Digital Practices in History and Ethnography IG</td>
</tr>
<tr>
<td>13.</td>
<td>Domain Repositories Interest Group</td>
</tr>
<tr>
<td>14.</td>
<td>Education and Training on handling of research data</td>
</tr>
<tr>
<td>15.</td>
<td>ELIXIR Bridging Force IG</td>
</tr>
<tr>
<td>16.</td>
<td>Engagement IG</td>
</tr>
<tr>
<td>17.</td>
<td>Federated Identity Management</td>
</tr>
<tr>
<td>18.</td>
<td>Geospatial IG</td>
</tr>
<tr>
<td>19.</td>
<td>Libraries for Research Data</td>
</tr>
<tr>
<td>20.</td>
<td>Long tail of research data IG</td>
</tr>
<tr>
<td>21.</td>
<td>Marine Data Harmonization IG</td>
</tr>
<tr>
<td>22.</td>
<td>Metabolomics</td>
</tr>
<tr>
<td>23.</td>
<td>Metadata IG</td>
</tr>
<tr>
<td>24.</td>
<td>PID Interest Group</td>
</tr>
<tr>
<td>25.</td>
<td>Preservation e-Infrastructure IG</td>
</tr>
<tr>
<td>26.</td>
<td>Quality of Urban Life Interest Group</td>
</tr>
<tr>
<td>27.</td>
<td>RDA/CODATA Legal Interoperability IG</td>
</tr>
<tr>
<td>28.</td>
<td>RDA/CODATA Materials Data, Infrastructure &amp; Interoperability IG</td>
</tr>
<tr>
<td>29.</td>
<td>RDA/WDS Certification of Digital Repositories IG</td>
</tr>
<tr>
<td>30.</td>
<td>RDA/WDS Publishing Data Cost Recovery for Data Centres</td>
</tr>
<tr>
<td>31.</td>
<td>RDA/WDS Publishing Data IG</td>
</tr>
<tr>
<td>32.</td>
<td>Reproducibility IG</td>
</tr>
<tr>
<td>33.</td>
<td>Research data needs of the Photon and Neutron Science community</td>
</tr>
<tr>
<td>34.</td>
<td>Research Data Provenance</td>
</tr>
<tr>
<td>35.</td>
<td>Service Management IG</td>
</tr>
<tr>
<td>36.</td>
<td>Structural Biology IG</td>
</tr>
<tr>
<td>37.</td>
<td>Toxicogenomics Interoperability IG</td>
</tr>
</tbody>
</table>

* in review
The Research Data Alliance Community Today

Total RDA Community Members: 3029
from 103 countries

North America 37%
South America 1%
Africa 3%
Asia 6%
Australasia 4%
Europe 49%

Academia
Government/Public service 459
Other 173
SMEs 162
IT Consultancy/Development 98
Large Enterprise 66
Policy/Funding Agency 45
Press & Media 19

research data sharing without barriers
rd-alliance.org
Some themes amidst the difference

1. **Persistent Identifiers** for data, documents, people, organisations, instruments—Everything!

2. **Certifying Trust** in assertions, evidence, organisations, processes…

3. The value of **Conversations, Relationships, and Mediation** — an agile network effect.
An Area of Convergence and Agreement

Internet Domain
nodes with IP numbers
packages being exchanged
standardized protocols
An Area of Convergence and Agreement

Internet Domain
- nodes with IP numbers
- packages being exchanged
- standardized protocols

Data Domain
- objects with PID numbers
- objects being exchanged
- standardized protocols

Slide courtesy P. Wittenberg from L. Lannom from D. Clark
Some themes amidst the difference

1. **Persistent Identifiers** for data, documents, people, organisations, instruments—Everything!

2. **Certifying Trust** in assertions, evidence, organisations, processes…

3. The value of **Conversations, Relationships, and Mediation** — an agile network effect.
Increasing Complexity of Mediation

From: C. Borgman, 2008, NSF Cyberlearning Report
An Agile Manifesto for Organisations
(courtesy Bruce Caron)

We value:

• **Individuals and interactions** over processes and tools
• **Working volunteers** over comprehensive documentation
• **Member collaboration** over contract negotiation
• **Responding to change** over following a plan.
Working Groups Clusters

Social dimension

Q4
- Repository Audit and Certification DSA–WDS Partnership
- Brokering Governance
- Standardisation of Data Categories and Codes

Q3
- PID Information Types
- Data Type Registries
- RDA/WDS Publishing Data Workflows
- RDA/WDS Publishing Data Bibliometrics

Q2
- RDA/WDS Publishing Data Services

Q1
- RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
- Metadata Standards Directory
- Data Description Registry Interoperability
- Wheat Data Interoperability

Technical dimension

Data providers

Beneficiary dimension

Data consumers
The “Data Fabric”

Slide courtesy Peter Wittenburg
The “Data Fabric”

- DFT: Data Foundation & Terminology
- MD: Metadata
- PIT: PID Information Types
- DTR: Data Type Registries
- PP: Practical Policy

Data fabric domain:
- collection
- documented processing
- processing
- permanent store
- citing

Publishing domain:
- data & paper publications

Raw data:
- registration

Slide courtesy Peter Wittenburg
The “Data Fabric”

- **DFT** - Data Foundation & Terminology
- **MD** - Metadata
- **PIT** - PID Information Types
- **DTR** - Data Type Registries
- **PP** - Practical Policy
- **DMP** - Data Management Plan
- **DOM** - Data Object Metadata
- **CERT** - Certificate

Slide courtesy Peter Wittenburg
The “Data Fabric”
The “Data Fabric”

Data fabric domain

documented processing

processing

collection

DFT
Data Foundation & Terminology
simple model for digital data in a registered domain

MD
Metadata
community curated standards catalogue for metadata interoperability

PIT
PID Information Types
common interface for providers and users of persistent ID services worldwide with harmonized categories

DTR
Data Type Registries
allowing humans and machines to act on unknown data types

PP
Practical Policy
defining best practice workflows how to deal with data automatically and in a documented way

registration

permanent store

citing

data & paper publications

raw data

collection

CITDD

DMP

REPRO

PROV

PP

BDA

BROK

CERT

REP

DOM

CERT

Slide courtesy Peter Wittenburg
The “Data Fabric”

Data fabric domain

documented processing

processing

collection

DFT
Data Foundation & Terminology
simple model for digital data in a registered domain

MD
Metadata
community curated standards catalogue for metadata interoperability

PIT
PID Information Types
common interface for providers and users of persistent ID services worldwide with harmonized categories

DTR
Data Type Registries
allowing humans and machines to act on unknown data types

PP
Practical Policy
defining best practice workflows how to deal with data automatically and in a documented way

DMP
DOM
CITDD
FIM
REP
CERT

Slide courtesy Peter Wittenburg
WG/IG Tags and Titles

- Only words that occur twice or more
- ‘Data’ and common words removed
- Only half the Groups have tags
Regional RDAs

- Australian National Data Service, RDA/United States, RDA/Europe,
- Implement RDA deliverables locally and enhance adoption.
- Ensure regional or national issues are addressed globally.
- Support plenaries and support attendance at plenaries.
Initial Impact

• Data are having their day! RDA is both cause and effect.

• Collaborative value
  • Accelerating harmonization—the citation story.
  • Discovering shared themes—PIDs, the “platform” story
  • New insight from rethinking old paradigms—roadmaps, architectures, and lifecycles are passé. Reuse, agility and bridging are hip.

• Real deliverables in 2 years.
  • Learning fast through openness and 18 months.
  • demonstration of delivery from outputs book

• Money
  • funding success is advancing the field
  • measurable return on investment

• http://datastories.jiscinvolve.org/wp/
Preservation of LHC data
100PB growing to ~5EB for decades

<table>
<thead>
<tr>
<th>Veni</th>
<th>... to RDA plenaries, WG &amp; IG meetings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vidi</td>
<td>... <em>how other disciplines attacked the problem</em></td>
</tr>
<tr>
<td>Vici</td>
<td>... developed, refined and now implementing a strategy, including cost model and business case</td>
</tr>
</tbody>
</table>

→ A better solution, more sustainable and advanced by years

slide courtesy Jamie Shiers, CERN
Next Steps for RDA: Stay Pragmatic, Focus on Impact

More Infrastructure
- Continuing pipeline of infrastructure deliverables adopted, used, coordinated and amplified to accelerate data sharing

More effective Community
- Increasing coordination and collaboration between domains, sectors, organizations, communities.
- Effective advocacy for national and international data issues and communities.

Impact-focused Outreach
- Stronger partnerships with industry, governments, domains, organizations.
- Substantive engagement of students and early career professionals, greater spectrum of international cultures.

Next Plenaries (Plenaries are both community and working meetings. Meetings held twice yearly around the world.):
- September, 2015: Paris, France (P6)
- March, 2016: Tokyo, Japan (P7)
- September, 2016: ~ Washington, DC (P8)
- March, 2017: Barcelona, Spain (P9)

Joining RDA:
Go to rd-alliance.org and register
- Must agree to RDA principles (openness, community-driven, etc.)
- Free for individuals

Fran Berman, Research Data Alliance
Plenary 6 and Data Challenge!

CNAM, Paris, France
23 - 25 September 2015
RDA Membership

Interest Groups
- domain coordination, idea generation, maintenance, ...

Working Groups
- implementable, impactful outcomes

Technical Advisory Board
- socio-technical vision and strategy

Secretariat
- administration and operations

Organisational Advisory Board
- needs, adoption, business advice

Council
- organisational vision and strategy

Funders Forum

RDA Foundation