Indiana University Research Technology and the Research Data Alliance

Rob Quick
Manager High Throughput Computing
Operations Officer - OSG and SWAMP
Board Member - RDA Organizational Assembly

- RDA Mission and History

- RDA Structure

- RDA Outputs

- IU RT Organizational Goals

The Research Data Alliance builds the social and technical bridges that enable open sharing of data.

The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.

FAIR Open Data Principles

- Findable

- Accessible

- Interoperable

- Re-usable

"Hey wait a minute! My Big Data analysis says I'm one gift short..."

A brief history of IU and the RDA

- 1st Plenary Meeting - March, 2013 - Launch

  - EC, NSF, NIST, and ANDS (Australia)

  - IU SoIC - Beth Plale - Technical Advisory Board

- 4th Plenary Meeting - September, 2014 - IU Representative attended the OAB

- 6th Plenary Meeting - September, 2015 - IU officially an organizational member

- Currently 3000 members from 102 countries

RDA Structure

- Council - Oversight

- Secretariat Staff - Administration

- Technical Advisory Board - Development, maintenance and evolution of the Technical Roadmap

- Organizational Advisory Board - Provides organizational advise to the council

Structure

	Interest Groups	Working Groups	BoFs
Membership	All RDA members can join	All RDA members can join	All RDA members can join
Duration	As long as the group is active	12-18 months	Short-term around a Plenary
Outcomes	Possibly case statements for new WGs, coordination, communication	Concrete deliverables - "Running code", tools, standards, etc.	Convert to IG, convert to WG, or retire BoF
RDA liaisons	1 Tab member, 1 Secretariat member	1-2 TAB members, 1 Secretariat member	1 Secretariat member
Review process	Community review (4+ weeks), then review by TAB (2 weeks) and Council (2 weeks)	Community review (4+ weeks), then review by TAB (4 weeks) and Council (2 weeks)	Review by TAB

Interest Groups

- Agriculture Data
- Active Data Management Plans
- Big Data
- Biodiversity Data Integration
- Brokering
- Chemistry Research Data
- Data Fabric
- Data for Development
- Data Foundations and Terminology
- Data in Context
- Data Rescue

- Development of cloud computing capacity in world research
- Digital Practices in History and Ethnography
- Domain Repositories
- Education and Training on handling of research data
- ELIXIR Bridging Force
- Engagement
- Ethics and Social Aspects
- Federated Identity Management
- Geospatial
- Libraries for Research Data

- Long Tail of Research Data
- Marine Data Harmonization
- Metabolomics Data Interoperability
- Metadata
- National Data Services
- PID
- Preservation e-Infrastructure
- Quality of Urban Life
- RDA/CODATA Legal Interoperability
- RDA/CODATA Materials Data, Infrastructure, and Interoperability

- RDA/WDS Certification of Digital Repositories
- RDA/WDS Publishing Data Cost Recovery for Data Centers
- Repository Platforms for Research Data
- Reproducibility
- Research data needs of the Photon and Neutron Science Community
- Research Data Provenance
- Service Management
- Structural Biology
- Vocabulary Services

Interest Groups

- Agriculture Data
- Active Data Management Plans
- Big Data
- Biodiversity Data Integration
- Brokering
- Chemistry Research Data
- Data Fabric
- Data for Development
- Data Foundations and Terminology
- Data in Context
- Data Rescue

- Development of cloud computing capacity in world research
- Digital Practices in History and Ethnography
- Domain Repositories
- Education and Training on handling of research data
- ELIXIR Bridging Force
- Engagement
- Ethics and Social Aspects
- Federated Identity Management
- Geospatial
- Libraries for Research Data

- Long Tail of Research Data
- Marine Data Harmonization
- Metabolomics Data Interoperability
- Metadata
- National Data Services
- PID
- Preservation e-Infrastructure
- Quality of Urban Life
- RDA/CODATA Legal Interoperability
- RDA/CODATA Materials Data, Infrastructure, and Interoperability

- RDA/WDS Certification of Digital Repositories
- RDA/WDS Publishing Data Cost Recovery for Data Centers
- Repository Platforms for Research Data
- Reproducibility
- Research data needs of the Photon and Neutron Science Community
- Research Data Provenance
- Service Management
- Structural Biology
- Vocabulary Services

Interest Groups

- Agriculture Data

- Active Data Management Plans

- Big Data

- Biodiversity Data Integration

- Brokering

- Chemistry Research Data

- Data Fabric

- Data for Development

- Data Foundations and Terminology

- Data in Context

- Data Rescue

- Development of cloud computing capacity in world research

- Digital Practices in History and Ethnography

- Domain Repositories

- Education and Training on handling of research data

- ELIXIR Bridging Force

- Engagement

- Ethics and Social Aspects

- Federated Identity Management

- Geospatial

- Libraries for Research Data

- Long Tail of Research Data

- Marine Data Harmonization

- Metabolomics Data Interoperability

- Metadata

- National Data Services

- PID

- Preservation e-Infrastructure

- Quality of Urban Life

- RDA/CODATA Legal Interoperability

- RDA/CODATA Materials Data, Infrastructure, and Interoperability

- RDA/WDS Certification of Digital Repositories

- RDA/WDS Publishing Data Cost Recovery for Data Centers

- Repository Platforms for Research Data

- Reproducibility

- Research data needs of the Photon and Neutron Science Community

- Research Data Provenance

- Service Management

- Structural Biology

- Vocabulary Services

Interest Groups

- Agriculture Data

- Active Data Management Plans

- Big Data

- Biodiversity Data Integration

- Brokering

- Chemistry Research Data

- Data Fabric

- Data for Development

- Data Foundations and Terminology

- Data in Context

- Data Rescue

- Development of cloud computing capacity in world research

- Digital Practices in History and Ethnography

- Domain Repositories

- Education and Training on handling of research data

- ELIXIR Bridging Force

- Engagement

- Ethics and Social Aspects

- Federated Identity Management

- Geospatial

- Libraries for Research Data

- Long Tail of Research Data

- Marine Data Harmonization

- Metabolomics Data Interoperability

- Metadata

- National Data Services

- PID

- Preservation e-Infrastructure

- Quality of Urban Life

- RDA/CODATA Legal Interoperability

- RDA/CODATA Materials Data, Infrastructure, and Interoperability

- RDA/WDS Certification of Digital Repositories

- RDA/WDS Publishing Data Cost Recovery for Data Centers

- Repository Platforms for Research Data

- Reproducibility

- Research data needs of the Photon and Neutron Science Community

- Research Data Provenance

- Service Management

- Structural Biology

- Vocabulary Services

Working Groups

- Array Database

- Brokering Governance

- Data Citation

- Data Description Registry Interoperability

- Data Foundations and Terminology

- Data Type Registries

- Metadata Standards Catalog

- Metadata Standards Directory

- PID Information Types

- Practical Policy

- QoS-DataLC Definitions

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

- RDA/WDS Publishing Data Bibliometrics

- RDA/WDS Publishing Data Services

- RDA/WDS Publishing Data Workflows

- Repository Audit and Certification DSA-WDS Partnership

- Research Data Collections

- The BioSharing Registry

- Wheat Data Interoperability

- Security and Trust

Working Group Outputs

Data Foundations and Terminology Working Group

- No common language between data communities

- At some level of abstraction, management and organization of data is independent of content

- 21 data models and 120 interviews

- Defined a number of simple definitions and

- http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

Data Type Registries Working Group

- Researchers can gather data from a wide variety of sources

- However, if the data is not typed it is often hard use and often discarded

- A data type registry allows data producers to record the implicit details of their data

- http://typeregistry.org

PID Information Types Working Group

- Several providers of Persistent Identifiers (PIDs)

- Thus serval different access methods (aka APIs)

- Checksums should be provided independently of the service provider

- Come up with a core set of information types

- Provide a common API

- Available as a prototype - documentation at https://www.rd-alliance.org/group/pid-information-types-wg.html

Practical Policy Working Group

- Data repositories require highly automated, safe, and documented management strategies

- With increased complexity published data procedures and policies are required to build trust in operations

- Collect typical policies, evaluate them, and provide best practice solutions

- "Policy Workbook" being created by DATASET and EUDAT

- https://www.rd-alliance.org/group/practical-policy-wg.html

Scalable Dynamic Data Citation Working Group

- Existing tools and data collections were not developed considering long term sustainability

- Big data sets are not used in their entirety and data sets are not static making a need for data identification and citation mechanisms that identify arbitrary subsets

- Devise a precise, machine actionable identification of arbitrary sub selections of data at a given point

- "RDA Recommendations for Data Citation of Evolving Data" - https://rd-alliance.org/rda-wgdc-recommendations-vers-sep-24-2015.html

Data Description Registry Interoperability Working Group

- Data repositories and registries are fragmented across institutions, countries and research domains

- Creating a series of bi-lateral information exchange projects

- Open, extensible, and flexible cross-platform research data discovery software solutions

- RD-Switchboard - uses research graph technology to link datasets - work is ongoing

Metadata Standards Directory

- A common challenge with research datasets is inconsistent approaches to in applying metadata or metadata schemes

- Create a metadata standard directory for users not automated tools and lay the foundation for a machine understandable catalogue of metadata standards

- http://rd-alliance.github.io/metadata-directory/

Wheat Data Interoperability Working Group

- To produce an adequate amount of wheat, worldwide yield increase must go from 1% to 1.6% a year

- To tackle this many researchers are doing experiments in fields and laboratories

- Data standard harmonization is a priority in promoting research

- Recommendations at http://ist.blogs.inra.fr/wdi/phenotypes/ and https://rd-alliance.org/groups/wheat-data-interoperability-wg.html

RDA/WDS Publishing Data Bibliometrics Working Group

- Statistical analysis of articles (bibliometrics) is essential to obtain measures of the quality and impact of research products

- Currently, it is not easy to apply these indicators to research outputs beyond academic journals

- Systems for bibliometrics do not scale and do not cope well with research data

- Data citation is getting better but still not ingrained in research communities

- We need a mechanism for evaluating the impact/use of datasets

Repository Audit and Certification DSA/WDS Partnership Working Group

- Certification standards and accreditation have been developed by DSA and WDS among others

- Both offer a basic certification standard for trustworthy digital repositories

- DSA focus on Humanities and Social Sciences, WDS focus on Earth and Space Science

- Plan to produce a harmonized catalog of criteria for certification of repositories along with a common set of procedures for repository assessment

- Proposal of a joint DSA-WDS Certification Procedure expected in Spring 2016

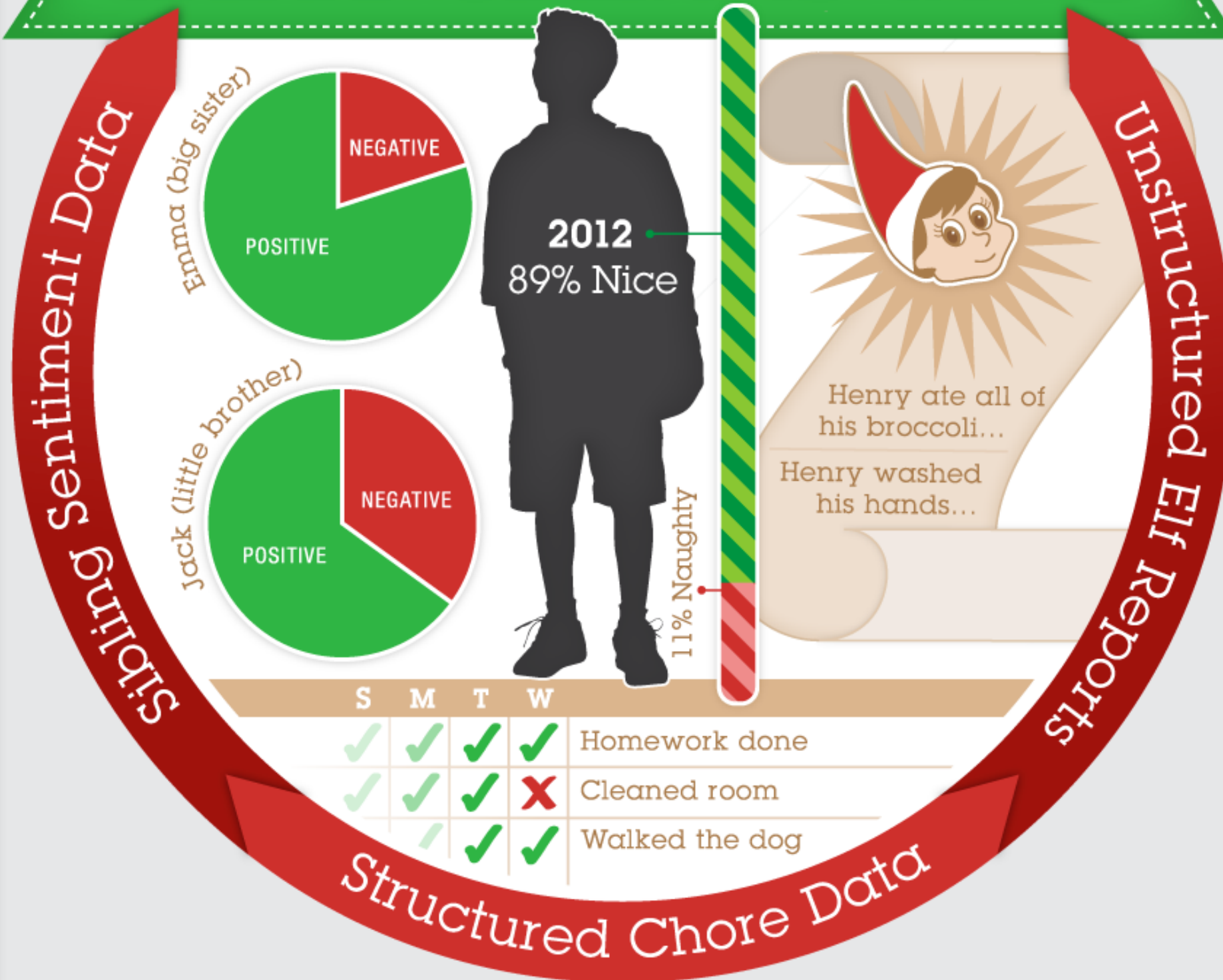# RDA/WDS Publishing Data Services Working Group

- No overarching standard to service to combine links between data sources and literature

- Bi-directional links provide visibility to both the publications and data along with context

- Building a universal interlinking service between data and scientific literature

- Prototype portal at http://dliservice.research-infrastructures.eu/ full fledged demo tool scheduled for Spring of 2016

RDA/WDS Publishing Data Workflows Working Group

- Researchers are encouraged to publish research data for reuse but there are insufficient incentives

- An analysis of data publishing models and solutions was needed to establish a solid future of data publishing workflows

- http://zenodo.org/record/20308#.Vlx63cqTrr8

- More analysis to come

OPERATION KRINGLE
THE BIG DATA PLATFORM OF THE NORTH POLE

Sibling Sentiment Data

Emma (big sister)
- NEGATIVE
- POSITIVE

Jack (little brother)
- NEGATIVE
- POSITIVE

2012
89% Nice

11% Naughty

Unstructured Elf Reports

Henry ate all of his broccoli...

Henry washed his hands...

	S	M	T	W	
Homework done	✓	✓	✓	✓	
Cleaned room	✓	✓	✓	✗	
Walked the dog	✓		✓	✓	

Structured Chore Data

Naughty & Nice Child Data (NNCD)

(NNCD × BILLIONS OF CHILDREN) EVERY DAY = BIG DATA!

IU RT Goals

- Promote data sharing and exchange, interoperability, use and re-use, discovery and analysis, stewardship and preservation

- Knowledge sharing

- Leads on important data projects and research opportunities

- Interactions with funding agencies and insight to what these funding agencies expect in Big Data proposals

- **Add value to IU Researcher Data!**

What does IU RT want/need from RDA?

- While there is participation in RDA from SoIC and Libraries, the organizational member is IU RT

- What would add value to the data we collect, analyze, curate, share?

- I'm prepared to sing carols, so I'd suggest providing feedback now

⭐

01001101

01100101 01110010

01110010 01111001 00100000

01000011 01101000 01110010 01101001

01110011 01110100 01101101 01100001 01110011

00100000 01100110 01110010 01101111 01101101 00100000

01010011 01110000 01101100 01110101

01101110

01101011