RESEARCH DATA ALLIANCE EUROPE

DK NATIONAL NODE

RDA groups and developments

Karsten Kryger Hansen

Aalborg University

AALBORG UNIVERSITY
DENMARK

Who am I?

- Data Management coordinator at AAU
- Involved in the DK-RDA national node
- Joined 5+ groups of RDA
  - Try to follow along
- Screen a number of RDA outputs
- Presented in a IG

Disclaimer

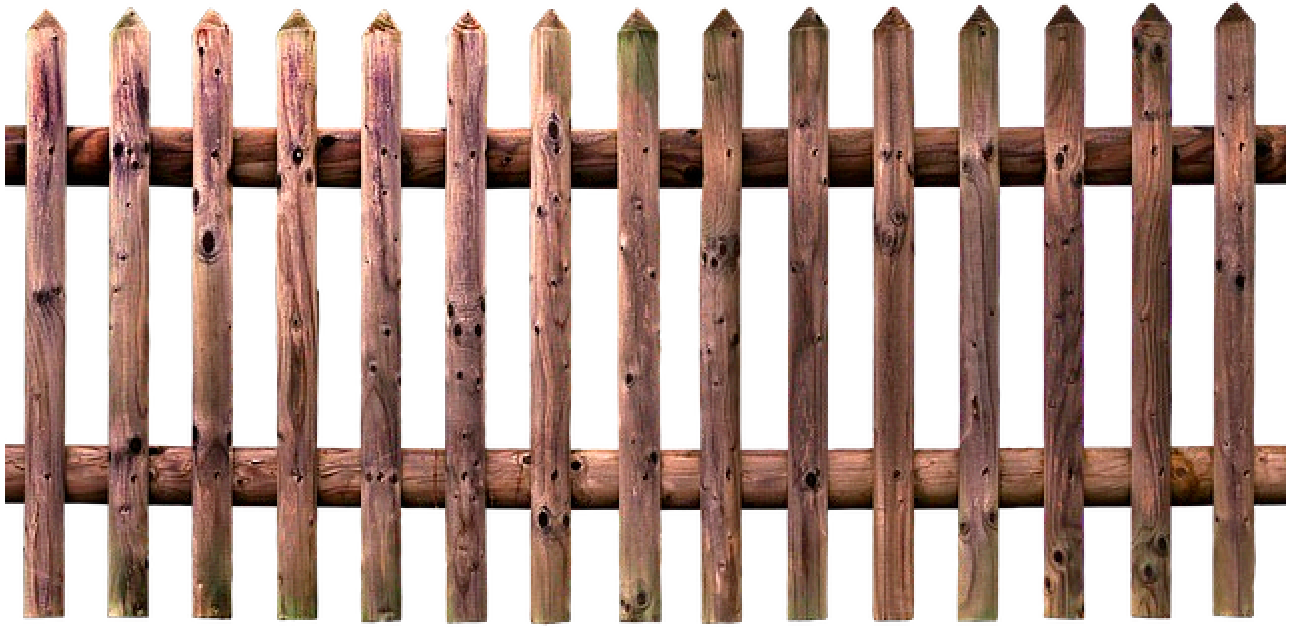
This presentation is not approved, nor adorsed by the RDA. All views are my personal views on RDA activities.

Navigating and making sense

The RDA group and recommendations landscape

Agenda for the next ~20 minutes

- Group structure
- Output and usage



https://pixabay.com/photos/fence-wood-fence-fence-element-3238491/

Group structures in RDA

Working Groups

- develop and implement tools, policy, practices and products for data management that are adopted and used by projects, organisations, communities
- TIMING: 12-18 MONTHS
- TOTAL: 36 Working Groups

RECOMMENDATIONS: Concrete deliverables - "Running code", tools, standards, etc.

Interest Groups

- focus on solving a specific data problem and identifying what kind of infrastructure needs to be built etc.
- TIMING: as long as group is active
- TOTAL: 66 Interest Groups

OUTPUT: Possibly case statements for new WGs, guidelines, best practice, etc.

+ BoF Groups

Working and interest groups

- Established by the community
  - WG: Case statements
  - IG: Charter
- Review process
  Community > TAB > Council.
- Co-chaired by community members
  2-4, at least 2 continents
- Associated liaison from RDA
- Explore synergy to other groups
- Meet at plenary sessions, and online
- Follow RDA Guiding Principles >>>

Openness – RDA community meetings and processes are open, and the deliverables of RDA Working Groups will be publicly disseminated.

Consensus – The RDA moves forward by achieving consensus among its membership. RDA processes and procedures include appropriate mechanisms to resolve conflicts.

Balance – The RDA seeks to promote balanced representation of its membership and stakeholder communities.

Harmonization – The RDA works to achieve harmonization across data standards, policies, technologies, infrastructure, and communities.

Community-driven – The RDA is a public, community-driven body constituted of volunteer members and organizations, supported by the RDA Secretariat.

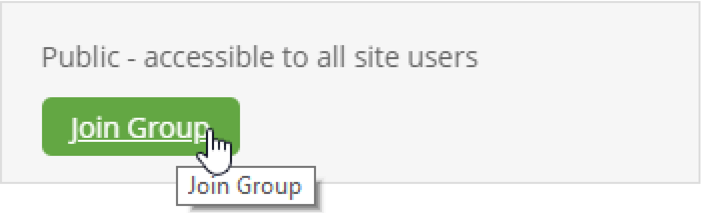
Non-profit - RDA does not promote, endorse, or sell commercial products, technologies, or services.

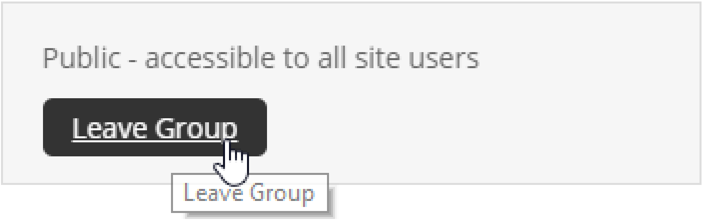
Working and Interest Groups (WG and IG)

- Free and easy to join
- No commitment

- Listen in
- Engage with what you can

- Catch up: Access to group history

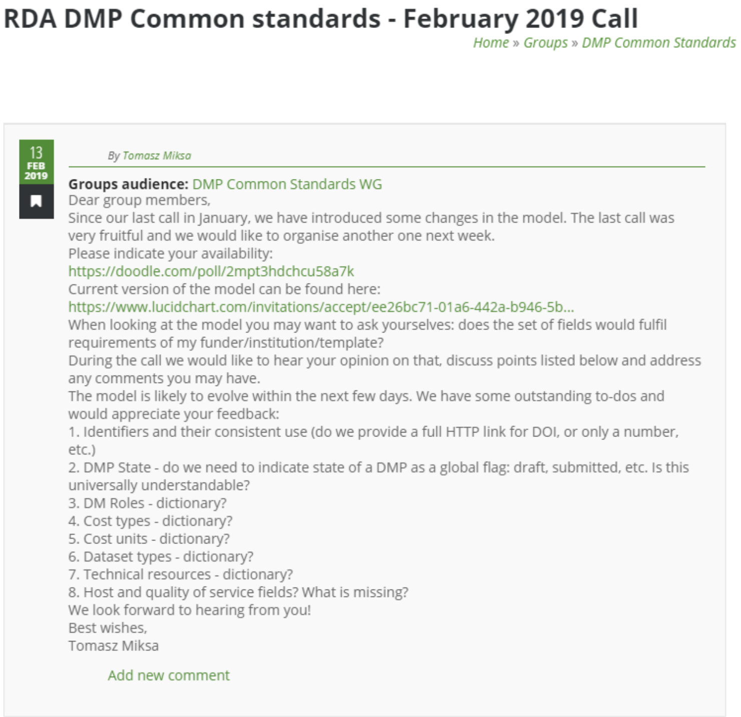

Public - accessible to all site users


Public - accessible to all site users

Working Groups

- Agrisemantics
- Array Database Assessment
- Big Data Modelling of UN SDGs
- Blockchain Applications in Health
- Brokering Framework Working Group
- Capacity Development for Agriculture Data
- Data Citation
- Data Description Registry Interoperability (DDRI)
- Data Type Registries & #2
- Data Usage Metrics
- Data Versioning
- DMP Common Standards
- Empirical Humanities Metadata Working Group
- Exposing Data Management Plans
- FAIR Data Maturity Model
- FAIRSharing Registry: connecting data policies, standards & databases
- Harmonizing FAIR descriptions of observational data
- International Materials Resource Registries
- Metadata Standards Catalog
- On-Farm Data Sharing (OFDS)
- Persistent Identification of Instruments
- PID Kernel Information
- Preserving Scientific Annotation
- Provenance Patterns
- RDA / TDWG Metadata Standards for attribution of physical and digital collections stewardship
- RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
- RDA/WDS Publishing Data Workflows
- RDA/WDS Scholarly Link Exchange (Scholix)
- Reproducible Health Data Services
- Research Data Collections
- Research Data Repository Interoperability
- Research Schemas
- Rice Data Interoperability
- Software Source Code Identification
- Storage Service Definitions
- WDS/RDA Assessment of Data Fitness for Use
- Wheat Data Interoperability

WG Example: WG DMP Common Standards



RDA DMP Common standards - February 2019 Call

Home » Groups » DMP Common Standards

13 FEB 2019

By Tomasz Miksa

Groups audience: DMP Common Standards WG

Dear group members,

Since our last call in January, we have introduced some changes in the model. The last call was very fruitful and we would like to organise another one next week.

Please indicate your availability:

https://doodle.com/poll/2mpt3hdchcu58a7k

Current version of the model can be found here:

https://www.lucidchart.com/invitations/accept/ee26bc71-01a6-442a-b946-5b...

When looking at the model you may want to ask yourselves: does the set of fields would fulfil requirements of my funder/institution/template?

During the call we would like to hear your opinion on that, discuss points listed below and address any comments you may have.

The model is likely to evolve within the next few days. We have some outstanding to-dos and would appreciate your feedback:

1. Identifiers and their consistent use (do we provide a full HTTP link for DOI, or only a number, etc.)
2. DMP State - do we need to indicate state of a DMP as a global flag: draft, submitted, etc. Is this universally understandable?
3. DM Roles - dictionary?
4. Cost types - dictionary?
5. Cost units - dictionary?
6. Dataset types - dictionary?
7. Technical resources - dictionary?
8. Host and quality of service fields? What is missing?

We look forward to hearing from you!

Best wishes,

Tomasz Miksa

Add new comment

Framing the scope of the common data model for machine-actionable Data Management Plans

Tomasz Miksa
SBA Research & TU Wien
Vienna, Austria
tmiksa@sba-research.org

João Cardoso
INESC-ID
& Instituto Superior Técnico
Lisbon, Portugal
joao.m.f.cardoso@tecnico.ulisboa.pt

José Borbinha
INESC-ID
& Instituto Superior Técnico
Lisbon, Portugal
jlb@tecnico.ulisboa.pt

Abstract—Currently, research requires processing data at a large scale. Data is not anymore a collection of static documents, but often a continuous stream of information flowing into information systems. Researchers need to manage their data efficiently not only to keep it safe, but also to ensure that it can be later correctly interpreted and reused. Existing solutions are not sufficient. Traditional Data Management Plans are manually created text documents that describe how research data will be handled. Yet, researchers must implement all actions by themselves. Machine-actionable Data Management Plans are a new approach that allows systems to act on behalf of researchers and other stakeholders involved in data management, to help them manage data in an efficient and scalable way. This paper summarises the results of work performed by the Research Data Alliance working group on Data Management Plan Common Standards to realise this vision. The paper describes results of consultations and proof of concept tools that help in: identifying needs for information of stakeholders involved in data management; defining the scope of the common data model for Machine-actionable Data Management Plans to allow for exchange of information between systems; identifying necessary services and components of infrastructure that support automation of data management tasks.

Index Terms—ata Management Plan, Machine-Actionable Data Management Plan, Workflowsata Management Plan, Machine-Actionable Data Management Plan, WorkflowsD

I. Introduction

With advances in technology, scientific research requires data processing in an increasingly larger scale. Data is no longer a collection of static documents, but often a continuous stream of information flowing into a repository [1], for example, satellite images or sensor data captured periodically. This new paradigm of research is often described as e-Science [2].

Researchers need to plan and manage their data efficiently, not only to keep it safe, but also to ensure that it can be later correctly interpreted and reused. This is especially important in the context of open data [3] and FAIR principles [4], [5].

One of the tools introduced to solve research Data Management (RDM) [6] challenges is the Data Management Plan (DMP) [7]. The overall objective of a DMP is to document, in a project, the techniques, methods and policies on how data is to be created, documented, accessed, preserved and disseminated. Various funding bodies, such as for example The National Science Foundation (NSF) or the European Commission (EC), already require that any funding application be accompanied by a DMP.

However, proper research data management, especially in view of big data and complex processing pipelines, is a complex task that requires cooperation of several stakeholders: not only researchers, but also, infrastructure operators, repository managers, legal experts, and so on. Researchers simply do not have enough expertise, nor time to prepare a DMP and then to actually implement it.

For this reason, there is a need for a solution that supports researchers in planning and managing data in an automated and scalable way. Research Data Alliance (RDA)[1] working group on DMP Common Standards[2] works to implement machine-actionable DMPs (maDMPs) [8]. The larger goal is to improve the experience for all involved by exchanging information across research tools and systems and embedding DMPs in existing workflows. As a result, parts of the DMP can be automatically generated and shared with other collaborators or funders. To achieve this goal there is the need for: good understanding of research data workflows, RDM infrastructure, common data model for maDMPs.

This paper presents the results to date of the RDA DMP Common Standards working group on realising maDMPs. It describes consultations performed and proof of concept tools developed that help in:

1) identifying stakeholders involved in data management and their requirements for information;
2) narrowing the scope of the common data model for maDMPs that acts as a standard for exchange of information between systems involved in data management;
3) identifying necessary services and components of infrastructure that support automation of data management tasks.

The paper is organised as follows. Section II provides definitions of the concepts of RDM, DMP, maDMP, and the RDA DMP Common Standards working group. Section III describes the work towards the creation of a DMP common model. Particular focus is given to the description of the two user consultations that were made to gather requirements for the development of the data model. In section IV we describe three tools that were developed as proof of concept, to demonstrate how a common DMP model can be used to automate tasks. Conclusions and outlook appear in section V.

II. Fundamentals

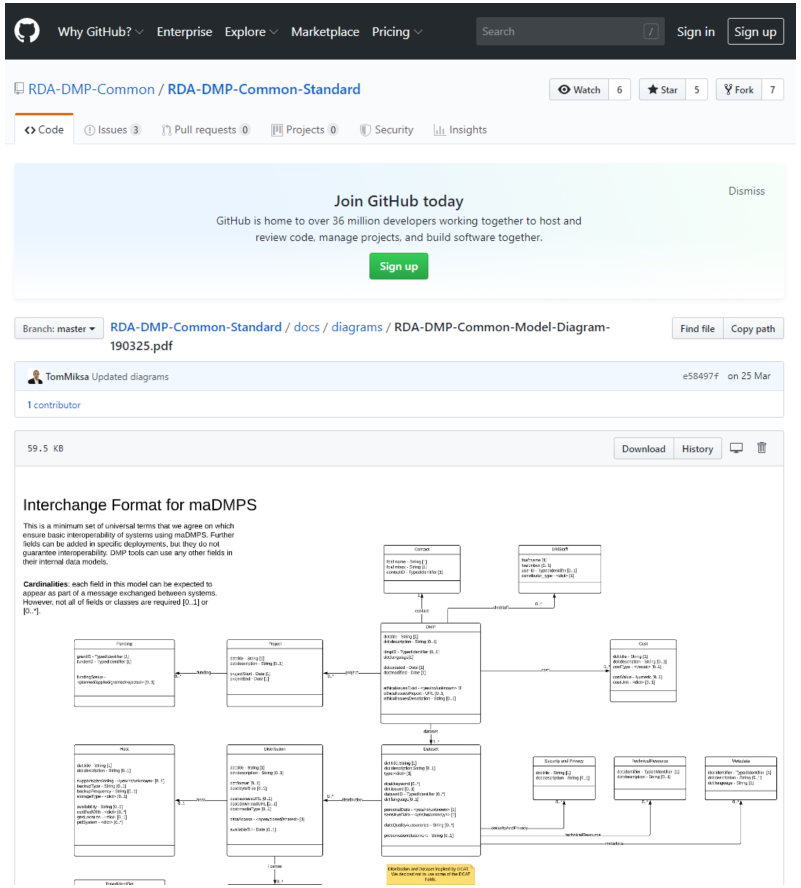
A. Research Data Management

As researchers must cope with the management of mounting quantities of data, and abide by the FAIR principles [4] of having data be findable, accessible, interoperable and reusable.

[1] https://www.rd-alliance.org

[2] https://www.rd-alliance.org/groups/dmp-common-standards-wg

https://doi.org/10.5281/zenodo.2161855



RDA-DMP-Common / RDA-DMP-Common-Standard

Join GitHub today

GitHub is home to over 36 million developers working together to host and review code, manage projects, and build software together.

Branch: master RDA-DMP-Common-Standard / docs / diagrams / RDA-DMP-Common-Model-Diagram-190325.pdf

TomMiksa Updated diagrams e58497f on 25 Mar

1 contributor

59.5 KB

Interchange Format for maDMPS

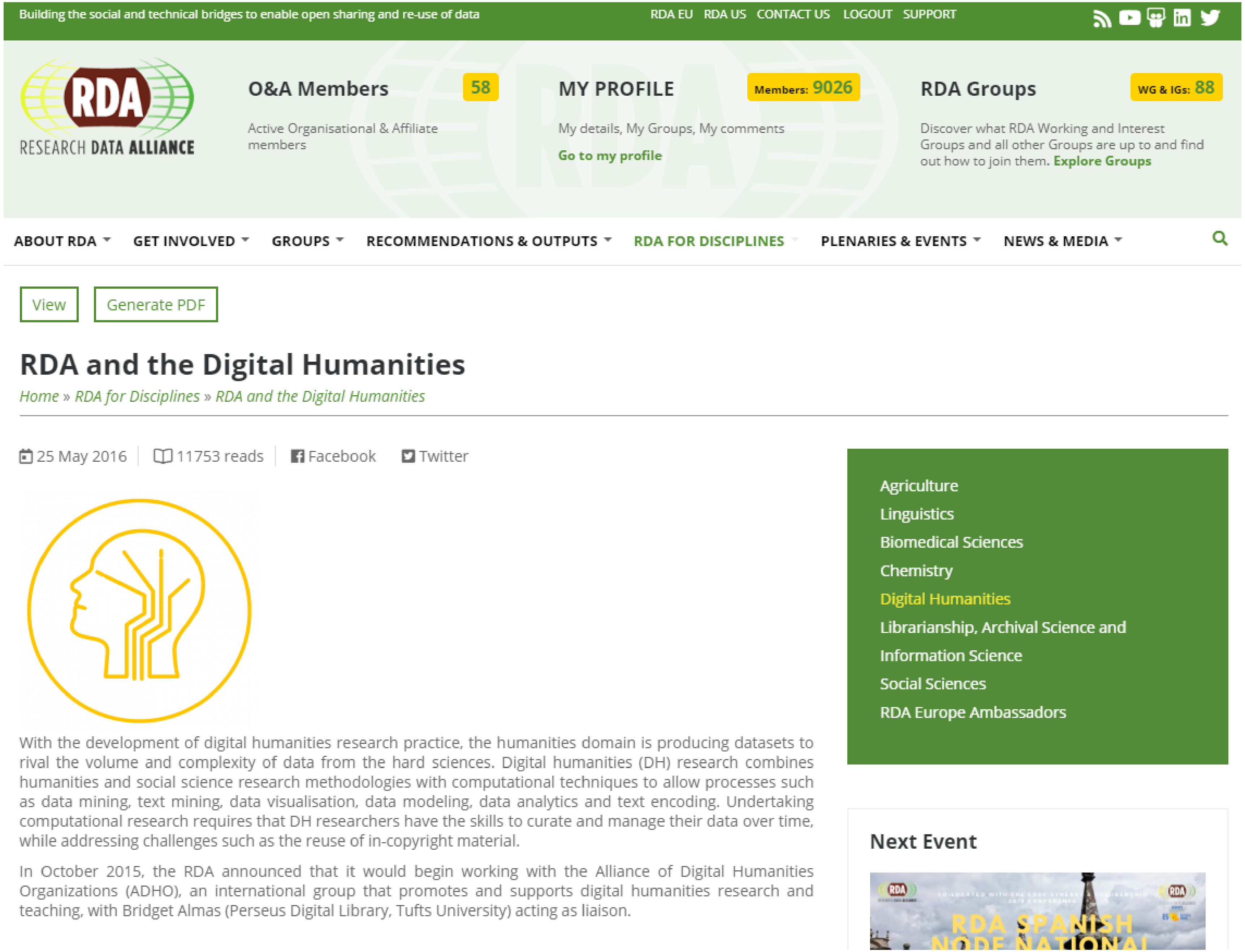
This is a minimum set of universal terms that we agree on which ensure basic interoperability of systems using maDMPS. Further fields can be added in specific deployments, but they do not guarantee interoperability. DMP tools can use any other fields in their internal data models.

Cardinalities: each field in this model can be expected to appear as part of a message exchanged between systems. However, not all of fields or classes are required [0..1] or [0..*].

Interest Groups

- Active Data Management Plans
- Agricultural Data Interest Group (IGAD)
- Archives and Records Professionals for Research Data
- Big Data
- Biodiversity Data Integration
- Brokering
- Chemistry Research Data
- CODATA/RDA Research Data Science Schools for Low and Middle Income Countries
- Data Discovery Paradigms
- Data Economics
- Data Fabric
- Data for Development
- Data Foundations and Terminology
- Data in Context
- Data policy standardisation and implementation
- Data Rescue
- Development of cloud computing capacity and education in developing world research
- Digital Practices in History and Ethnography
- Disciplinary Collaboration Framework
- Domain Repositories
- Early Career and Engagement
- Education and Training on handling of research data
- ELIXIR Bridging Force
- ESIP/RDA Earth, Space, and Environmental Sciences
- Ethics and Social Aspects of Data
- Federated Identity Management
- From Observational Data to Information
- Geospatial
- Global Water Information
- GO FAIR
- Health Data
- IG for Surveying Open Data Practices
- International Indigenous Data Sovereignty
- Libraries for Research Data
- Linguistics Data
- Long tail of research data
- Marine Data Harmonization
- Metadata
- …

RESEARCH DATA ALLIANCE EUROPE
DK NATIONAL NODE



Building the social and technical bridges to enable open sharing and re-use of data

RDA EU RDA US CONTACT US LOGOUT SUPPORT

RESEARCH DATA ALLIANCE

O&A Members 58

Active Organisational & Affiliate members

MY PROFILE Members: 9026

My details, My Groups, My comments

Go to my profile

RDA Groups WG & IGs: 88

Discover what RDA Working and Interest Groups and all other Groups are up to and find out how to join them. Explore Groups

RDA and the Digital Humanities

Home » RDA for Disciplines » RDA and the Digital Humanities

25 May 2016 | 11753 reads | Facebook | Twitter

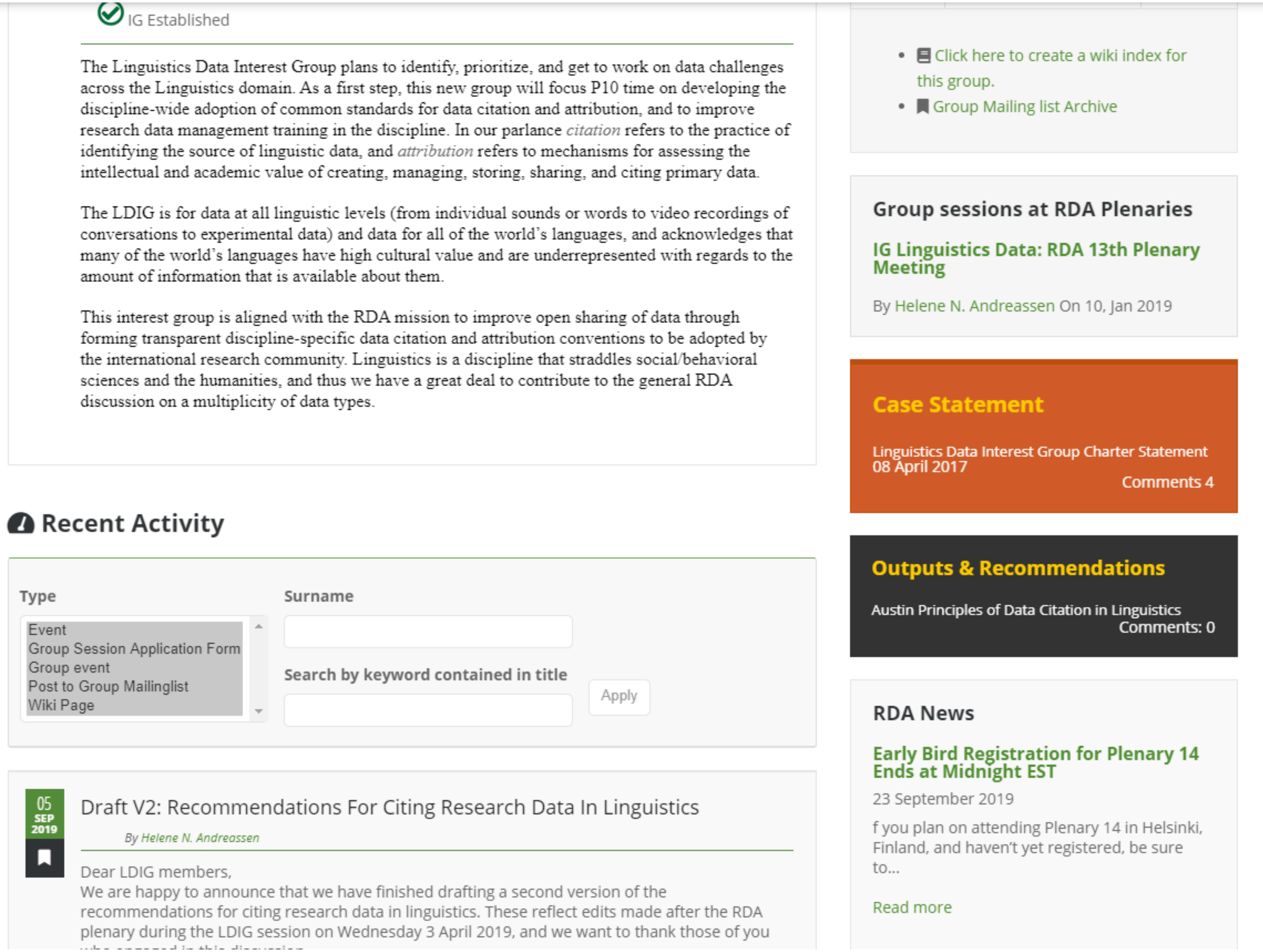
With the development of digital humanities research practice, the humanities domain is producing datasets to rival the volume and complexity of data from the hard sciences. Digital humanities (DH) research combines humanities and social science research methodologies with computational techniques to allow processes such as data mining, text mining, data visualisation, data modeling, data analytics and text encoding. Undertaking computational research requires that DH researchers have the skills to curate and manage their data over time, while addressing challenges such as the reuse of in-copyright material.

In October 2015, the RDA announced that it would begin working with the Alliance of Digital Humanities Organizations (ADHO), an international group that promotes and supports digital humanities research and teaching, with Bridget Almas (Perseus Digital Library, Tufts University) acting as liaison.

- Agriculture
- Linguistics
- Biomedical Sciences
- Chemistry
- Digital Humanities
- Librarianship, Archival Science and Information Science
- Social Sciences
- RDA Europe Ambassadors

Next Event

IG example #1: Linguistics Data IG



✓ IG Established

The Linguistics Data Interest Group plans to identify, prioritize, and get to work on data challenges across the Linguistics domain. As a first step, this new group will focus P10 time on developing the discipline-wide adoption of common standards for data citation and attribution, and to improve research data management training in the discipline. In our parlance *citation* refers to the practice of identifying the source of linguistic data, and *attribution* refers to mechanisms for assessing the intellectual and academic value of creating, managing, storing, sharing, and citing primary data.

The LDIG is for data at all linguistic levels (from individual sounds or words to video recordings of conversations to experimental data) and data for all of the world's languages, and acknowledges that many of the world's languages have high cultural value and are underrepresented with regards to the amount of information that is available about them.

This interest group is aligned with the RDA mission to improve open sharing of data through forming transparent discipline-specific data citation and attribution conventions to be adopted by the international research community. Linguistics is a discipline that straddles social/behavioral sciences and the humanities, and thus we have a great deal to contribute to the general RDA discussion on a multiplicity of data types.

Recent Activity

Type

Event
Group Session Application Form
Group event
Post to Group Mailinglist
Wiki Page

Surname

Search by keyword contained in title

Apply

05 SEP 2019

Draft V2: Recommendations For Citing Research Data In Linguistics

By Helene N. Andreassen

Dear LDIG members,
We are happy to announce that we have finished drafting a second version of the recommendations for citing research data in linguistics. These reflect edits made after the RDA plenary during the LDIG session on Wednesday 3 April 2019, and we want to thank those of you

- Click here to create a wiki index for this group.
- Group Mailing list Archive

Group sessions at RDA Plenaries

IG Linguistics Data: RDA 13th Plenary Meeting

By Helene N. Andreassen On 10, Jan 2019

Case Statement

Linguistics Data Interest Group Charter Statement
08 April 2017
Comments 4

Outputs & Recommendations

Austin Principles of Data Citation in Linguistics
Comments: 0

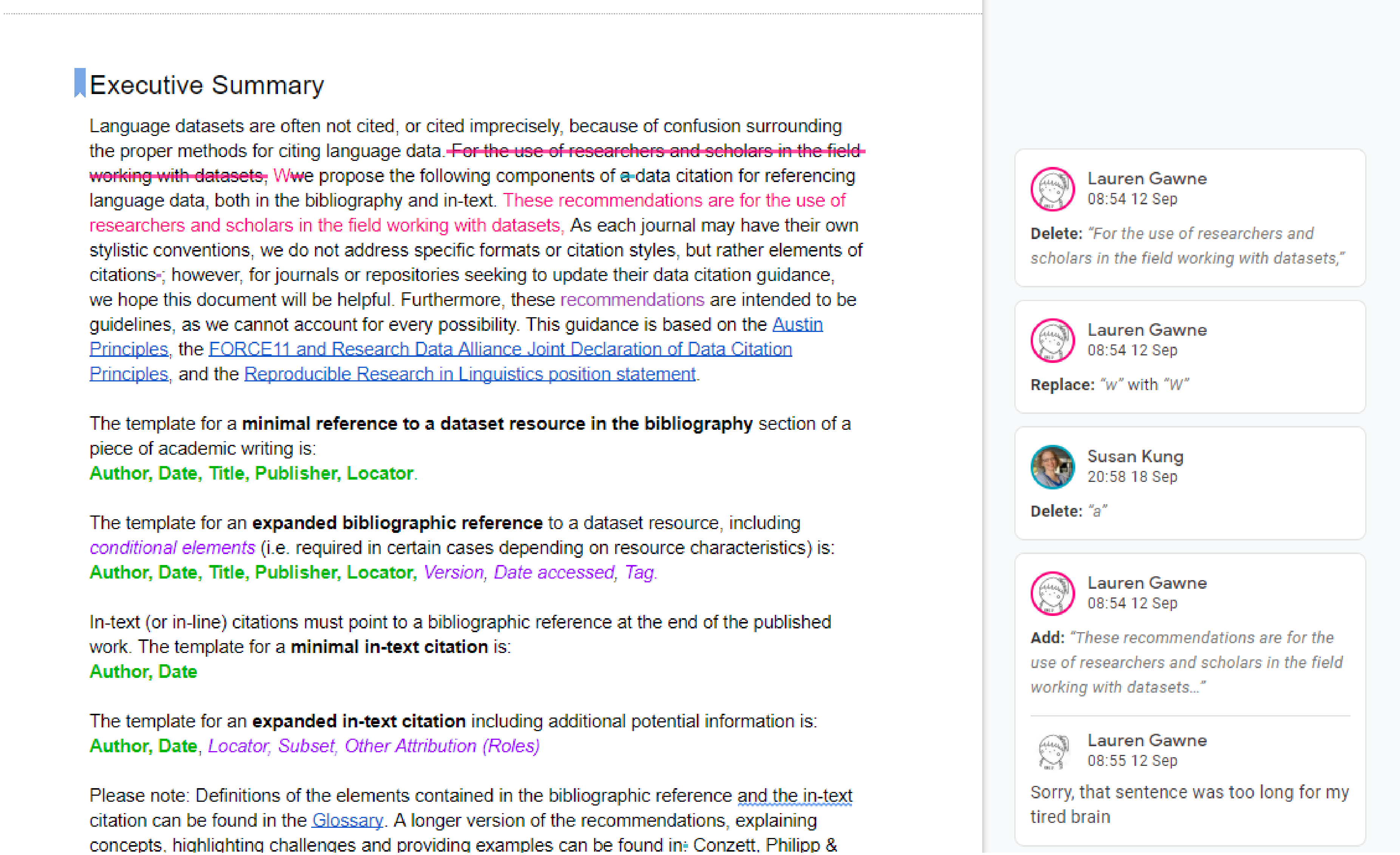
RDA News

Early Bird Registration for Plenary 14 Ends at Midnight EST

23 September 2019

f you plan on attending Plenary 14 in Helsinki, Finland, and haven't yet registered, be sure to...

Read more



Executive Summary

Language datasets are often not cited, or cited imprecisely, because of confusion surrounding the proper methods for citing language data. ~~For the use of researchers and scholars in the field working with datasets,~~ W~~w~~e propose the following components of ~~a~~ data citation for referencing language data, both in the bibliography and in-text. These recommendations are for the use of researchers and scholars in the field working with datasets, As each journal may have their own stylistic conventions, we do not address specific formats or citation styles, but rather elements of citations~~.~~; however, for journals or repositories seeking to update their data citation guidance, we hope this document will be helpful. Furthermore, these recommendations are intended to be guidelines, as we cannot account for every possibility. This guidance is based on the Austin Principles, the FORCE11 and Research Data Alliance Joint Declaration of Data Citation Principles, and the Reproducible Research in Linguistics position statement.

The template for a **minimal reference to a dataset resource in the bibliography** section of a piece of academic writing is:
Author, Date, Title, Publisher, Locator.

The template for an **expanded bibliographic reference** to a dataset resource, including *conditional elements* (i.e. required in certain cases depending on resource characteristics) is:
Author, Date, Title, Publisher, Locator, *Version, Date accessed, Tag.*

In-text (or in-line) citations must point to a bibliographic reference at the end of the published work. The template for a **minimal in-text citation** is:
Author, Date

The template for an **expanded in-text citation** including additional potential information is:
Author, Date, *Locator, Subset, Other Attribution (Roles)*

Please note: Definitions of the elements contained in the bibliographic reference and the in-text citation can be found in the Glossary. A longer version of the recommendations, explaining concepts, highlighting challenges and providing examples can be found in: Conzett, Philipp &

Lauren Gawne
08:54 12 Sep
Delete: *"For the use of researchers and scholars in the field working with datasets,"*

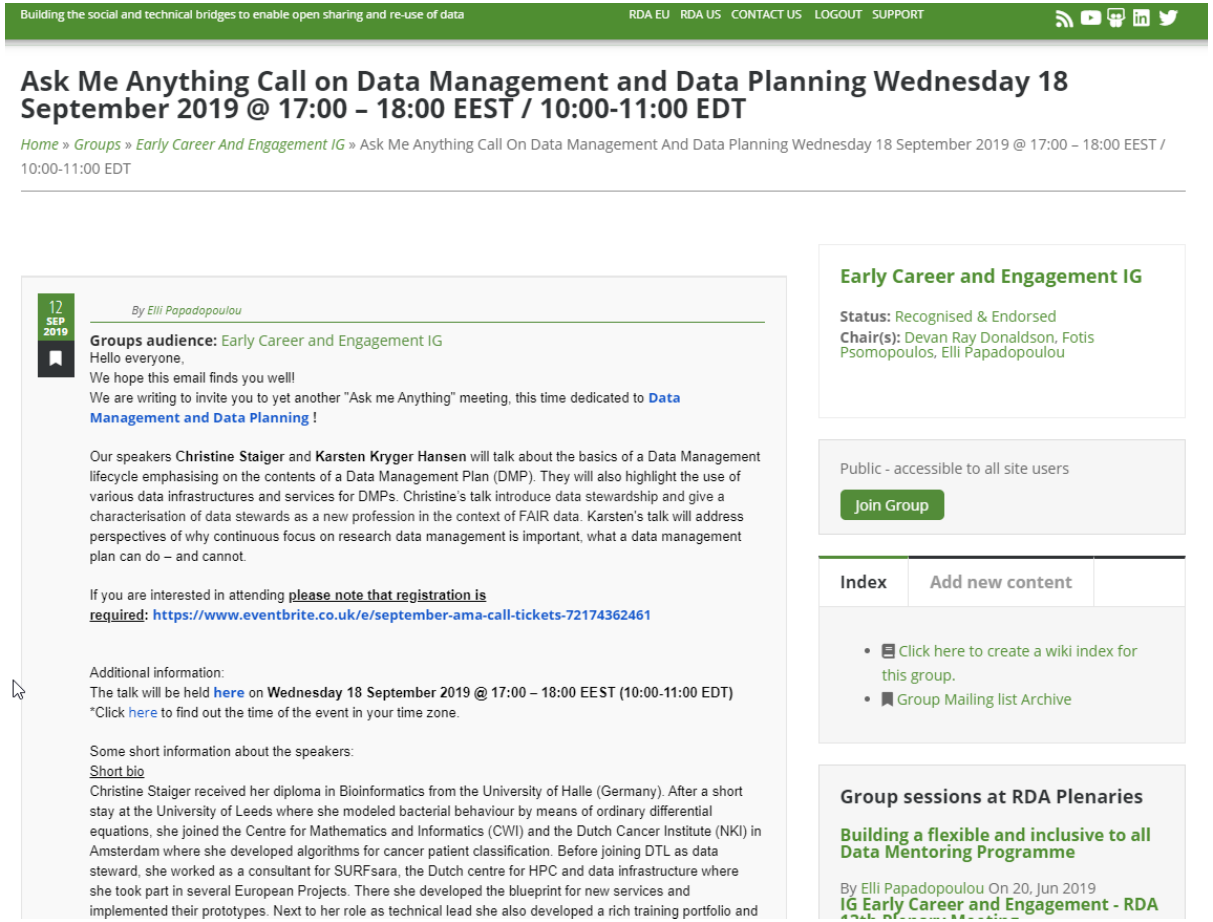
Lauren Gawne
08:54 12 Sep
Replace: *"w"* with *"W"*

Susan Kung
20:58 18 Sep
Delete: *"a"*

Lauren Gawne
08:54 12 Sep
Add: *"These recommendations are for the use of researchers and scholars in the field working with datasets…"*

Lauren Gawne
08:55 12 Sep
Sorry, that sentence was too long for my tired brain

IG example #2: Early Career and Engagement IG


Building the social and technical bridges to enable open sharing and re-use of data
RDA EU RDA US CONTACT US LOGOUT SUPPORT

Ask Me Anything Call on Data Management and Data Planning Wednesday 18 September 2019 @ 17:00 – 18:00 EEST / 10:00-11:00 EDT

Home » Groups » Early Career And Engagement IG » Ask Me Anything Call On Data Management And Data Planning Wednesday 18 September 2019 @ 17:00 – 18:00 EEST / 10:00-11:00 EDT

12 SEP 2019

By Elli Papadopoulou

Groups audience: Early Career and Engagement IG

Hello everyone,
We hope this email finds you well!
We are writing to invite you to yet another "Ask me Anything" meeting, this time dedicated to **Data Management and Data Planning** !

Our speakers **Christine Staiger** and **Karsten Kryger Hansen** will talk about the basics of a Data Management lifecycle emphasising on the contents of a Data Management Plan (DMP). They will also highlight the use of various data infrastructures and services for DMPs. Christine's talk introduce data stewardship and give a characterisation of data stewards as a new profession in the context of FAIR data. Karsten's talk will address perspectives of why continuous focus on research data management is important, what a data management plan can do – and cannot.

If you are interested in attending **please note that registration is required**: **https://www.eventbrite.co.uk/e/september-ama-call-tickets-72174362461**

Additional information:
The talk will be held here on **Wednesday 18 September 2019 @ 17:00 – 18:00 EEST (10:00-11:00 EDT)**
*Click here to find out the time of the event in your time zone.

Some short information about the speakers:
Short bio
Christine Staiger received her diploma in Bioinformatics from the University of Halle (Germany). After a short stay at the University of Leeds where she modeled bacterial behaviour by means of ordinary differential equations, she joined the Centre for Mathematics and Informatics (CWI) and the Dutch Cancer Institute (NKI) in Amsterdam where she developed algorithms for cancer patient classification. Before joining DTL as data steward, she worked as a consultant for SURFsara, the Dutch centre for HPC and data infrastructure where she took part in several European Projects. There she developed the blueprint for new services and implemented their prototypes. Next to her role as technical lead she also developed a rich training portfolio and

Early Career and Engagement IG

Status: Recognised & Endorsed
Chair(s): Devan Ray Donaldson, Fotis Psomopoulos, Elli Papadopoulou

Public - accessible to all site users

Join Group

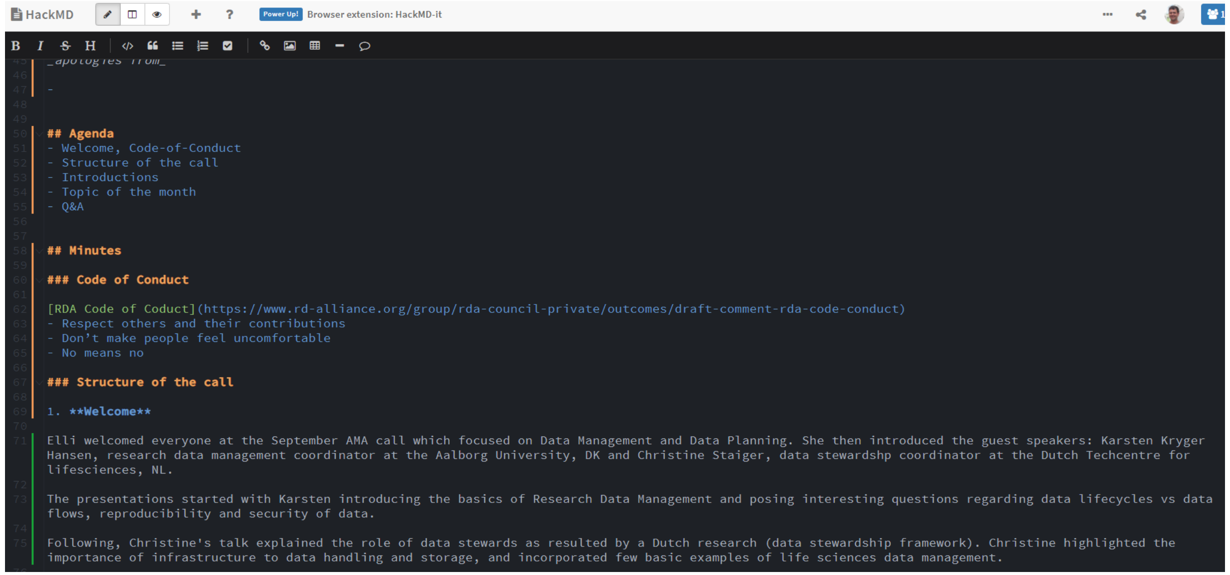
Index | **Add new content**

- Click here to create a wiki index for this group.
- Group Mailing list Archive

Group sessions at RDA Plenaries

Building a flexible and inclusive to all Data Mentoring Programme

By Elli Papadopoulou On 20, Jun 2019
IG Early Career and Engagement - RDA

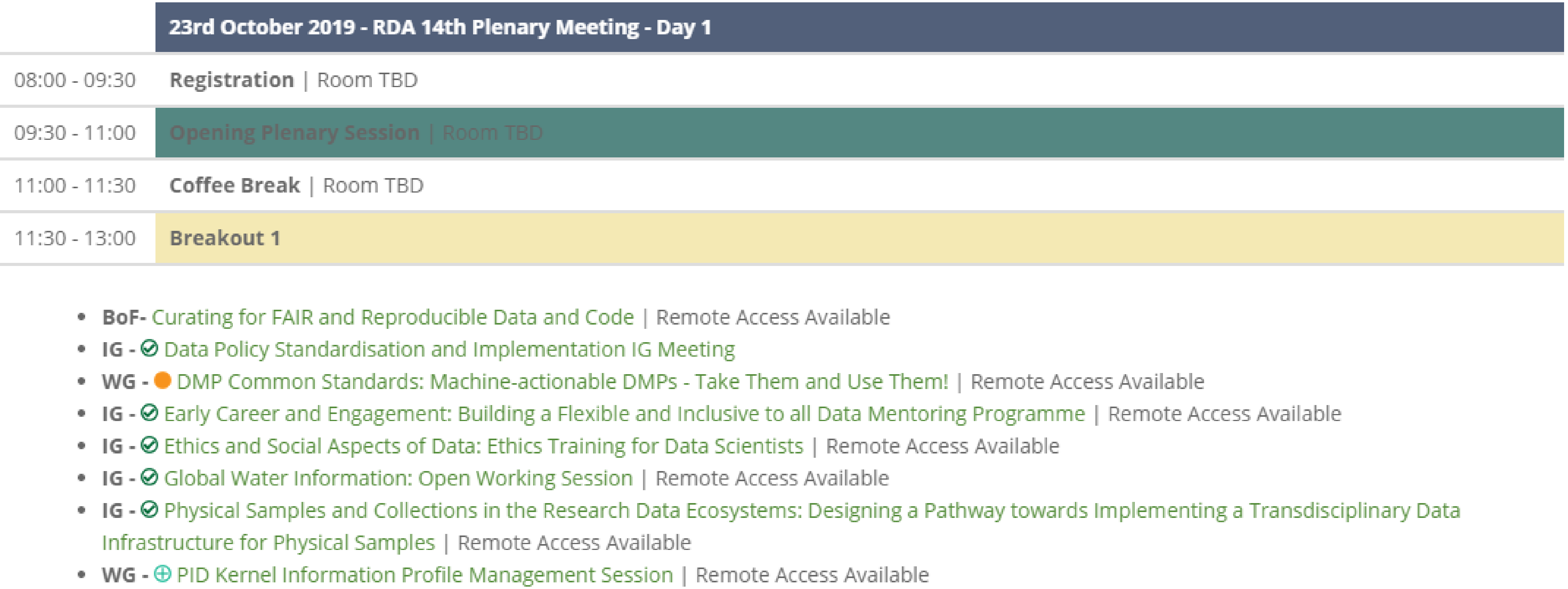


```
_apologies from_

-


## Agenda
- Welcome, Code-of-Conduct
- Structure of the call
- Introductions
- Topic of the month
- Q&A


## Minutes

### Code of Conduct

[RDA Code of Coduct](https://www.rd-alliance.org/group/rda-council-private/outcomes/draft-comment-rda-code-conduct)
- Respect others and their contributions
- Don't make people feel uncomfortable
- No means no

### Structure of the call

1. **Welcome**

Elli welcomed everyone at the September AMA call which focused on Data Management and Data Planning. She then introduced the guest speakers: Karsten Kryger Hansen, research data management coordinator at the Aalborg University, DK and Christine Staiger, data stewardshp coordinator at the Dutch Techcentre for lifesciences, NL.

The presentations started with Karsten introducing the basics of Research Data Management and posing interesting questions regarding data lifecycles vs data flows, reproducibility and security of data.

Following, Christine's talk explained the role of data stewards as resulted by a Dutch research (data stewardship framework). Christine highlighted the importance of infrastructure to data handling and storage, and incorporated few basic examples of life sciences data management.
```

RDA 14th Plenary Meeting

Helsinki, Finland - 23 - 25 October 2019

Early Bird Registration open until 24 September 2019 | Call for sponsors



	23rd October 2019 - RDA 14th Plenary Meeting - Day 1
08:00 - 09:30	**Registration** | Room TBD
09:30 - 11:00	**Opening Plenary Session** | Room TBD
11:00 - 11:30	**Coffee Break** | Room TBD
11:30 - 13:00	**Breakout 1**

- **BoF-** Curating for FAIR and Reproducible Data and Code | Remote Access Available
- **IG -** ✓ Data Policy Standardisation and Implementation IG Meeting
- **WG -** ● DMP Common Standards: Machine-actionable DMPs - Take Them and Use Them! | Remote Access Available
- **IG -** ✓ Early Career and Engagement: Building a Flexible and Inclusive to all Data Mentoring Programme | Remote Access Available
- **IG -** ✓ Ethics and Social Aspects of Data: Ethics Training for Data Scientists | Remote Access Available
- **IG -** ✓ Global Water Information: Open Working Session | Remote Access Available
- **IG -** ✓ Physical Samples and Collections in the Research Data Ecosystems: Designing a Pathway towards Implementing a Transdisciplinary Data Infrastructure for Physical Samples | Remote Access Available
- **WG -** ⊕ PID Kernel Information Profile Management Session | Remote Access Available

Outputs from RDA

Types of output

- Recommendations
  - The recommendation itself (reviewed)
  - Supporting outputs (reviewed)
  - Other outputs (not reviewed)
- Adoption use cases
- Adoption stories
- Standards

Use of outputs

- Technical specification supplement
- Procedure description
- Inspiration
- Network building

Recommendation workflow

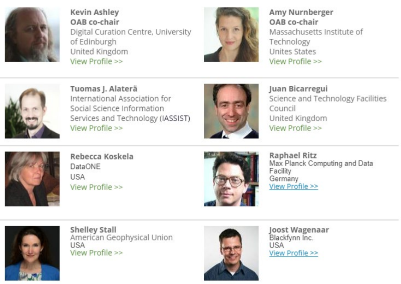
Request for comments ongoing!

Results of an Analysis of Existing FAIR Assessment Tools

WG delivers output to RDA secretariat

Community review (RfC) for 1 month

RDA Organisational Advisory Board review



Kevin Ashley
OAB co-chair
Digital Curation Centre, University of Edinburgh
United Kingdom
View Profile >>

Amy Nurnberger
OAB co-chair
Massachusetts Institute of Technology
Unites States
View Profile >>

Tuomas J. Alaterä
International Association for Social Science Information Services and Technology (IASSIST)
View Profile >>

Juan Bicarregui
Science and Technology Facilities Council
United Kingdom
View Profile >>

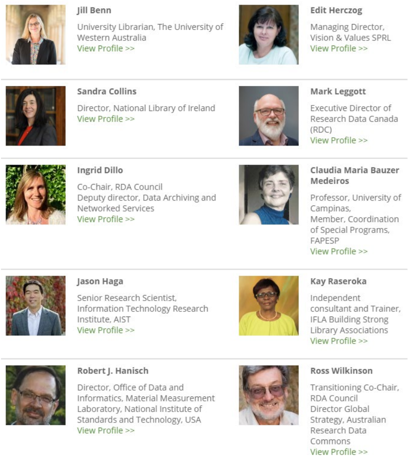
Rebecca Koskela
DataONE
USA
View Profile >>

Raphael Ritz
Max Planck Computing and Data Facility
Germany
View Profile >>

Shelley Stall
American Geophysical Union
USA
View Profile >>

Joost Wagenaar
Blackfynn Inc.
USA
View Profile >>

Council review and approval



Jill Benn
University Librarian, The University of Western Australia
View Profile >>

Edit Herczog
Managing Director, Vision & Values SPRL
View Profile >>

Sandra Collins
Director, National Library of Ireland
View Profile >>

Mark Leggott
Executive Director of Research Data Canada (RDC)
View Profile >>

Ingrid Dillo
Co-Chair, RDA Council
Deputy director, Data Archiving and Networked Services
View Profile >>

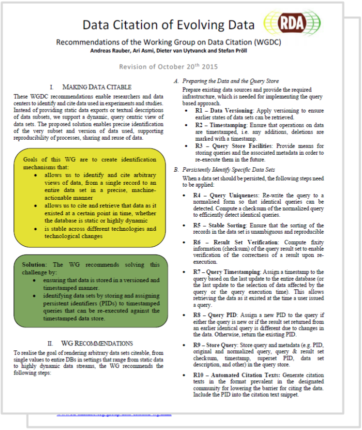
Claudia Maria Bauzer Medeiros
Professor, University of Campinas,
Member, Coordination of Special Programs, FAPESP
View Profile >>

Jason Haga
Senior Research Scientist, Information Technology Research Institute, AIST
View Profile >>

Kay Raseroka
Independent consultant and Trainer, IFLA Building Strong Library Associations
View Profile >>

Robert J. Hanisch
Director, Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, USA
View Profile >>

Ross Wilkinson
Transitioning Co-Chair, RDA Council
Director Global Strategy, Australian Research Data Commons
View Profile >>



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)
Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Revision of October 20th 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 subset 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

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.
- **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 Facilities**: Provide means for storing queries and the associated metadata in order to re-execute them in the future.

B. Persistently Identify Specific Data Sets

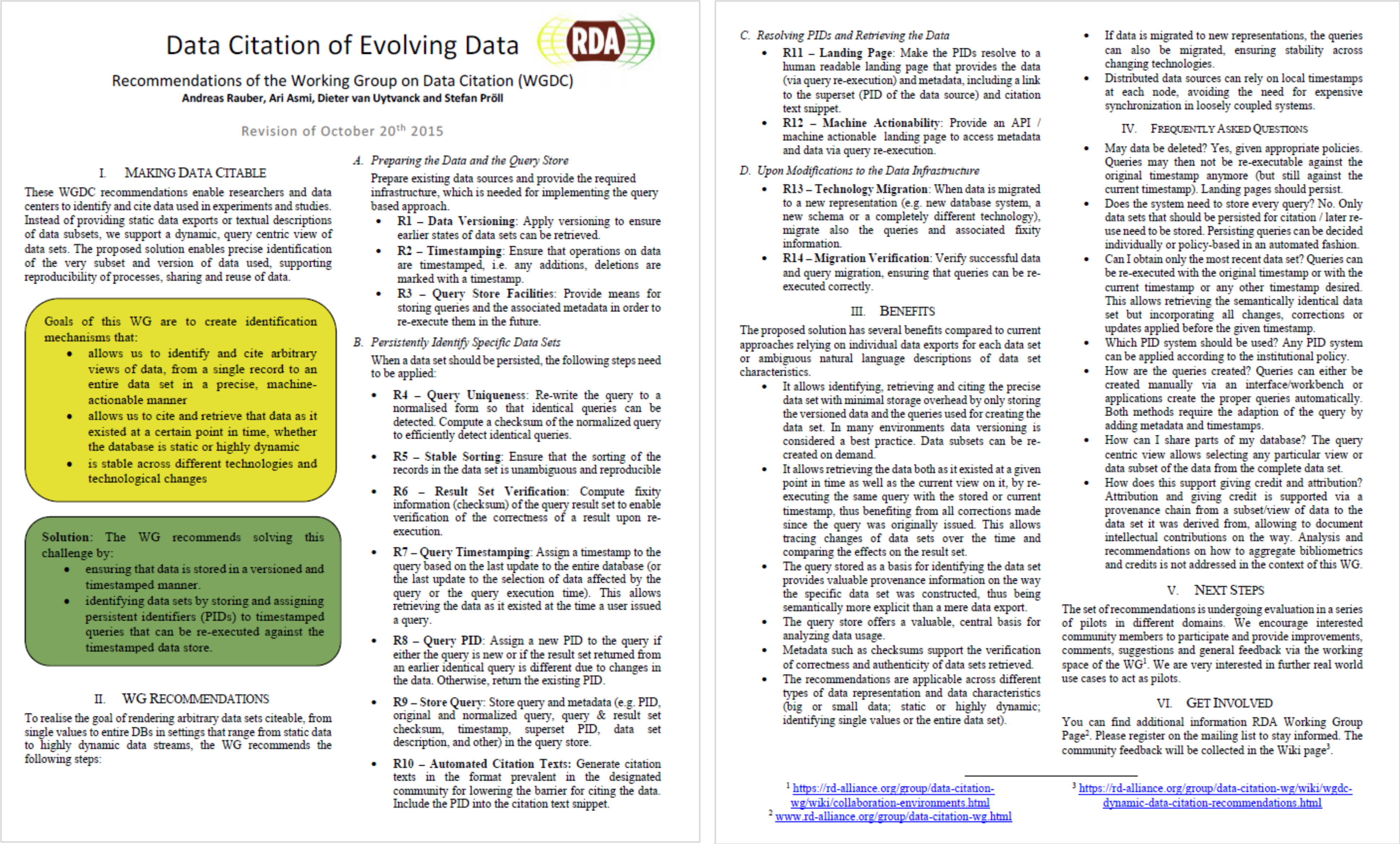
When a data set should be persisted, the following steps need to be applied:
- **R4 – Query Uniqueness**: Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.
- **R5 – Stable Sorting**: Ensure that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification**: Compute fixity information (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 the time a user issued a query.
- **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 normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.
- **R10 – Automated Citation Texts:** Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.

WG Pilot adoptions

Possible ICT standard

Example of a recommendation

Citation and Download: Andreas Rauber; Ari Asmi; Dieter van Uytvanck; Stefan Proell (2015): Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). DOI: 10.15497/RDA00016



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

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

Revision of October 20th 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 subset 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

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **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 Facilities**: Provide means for storing queries and the associated metadata in order to re-execute them in the future.

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 that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification**: Compute fixity information (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 the time a user issued a query.
- **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 normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.
- **R10 – Automated Citation Texts:** Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.

C. Resolving PIDs and Retrieving the Data

- **R11 – Landing Page**: Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.
- **R12 – Machine Actionability**: Provide an API / machine actionable landing page to access metadata and data via query re-execution.

D. Upon Modifications to the Data Infrastructure

- **R13 – Technology Migration**: When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated fixity information.
- **R14 – Migration Verification**: Verify successful data and query migration, ensuring that queries can be re-executed correctly.

III. Benefits

The proposed solution has several benefits compared to current approaches relying on individual data exports for each data set or ambiguous natural language descriptions of data set characteristics.

- It allows identifying, retrieving and citing the precise data set with minimal storage overhead by only storing the versioned data and the queries used for creating the data set. In many environments data versioning is considered a best practice. Data subsets can be re-created on demand.
- It allows retrieving the data both as it existed at a given point in time as well as the current view on it, by re-executing the same query with the stored or current timestamp, thus benefiting from all corrections made since the query was originally issued. This allows tracing changes of data sets over the time and comparing the effects on the result set.
- The query stored as a basis for identifying the data set provides valuable provenance information on the way the specific data set was constructed, thus being semantically more explicit than a mere data export.
- The query store offers a valuable, central basis for analyzing data usage.
- Metadata such as checksums support the verification of correctness and authenticity of data sets retrieved.
- The recommendations are applicable across different types of data representation and data characteristics (big or small data; static or highly dynamic; identifying single values or the entire data set).
- If data is migrated to new representations, the queries can also be migrated, ensuring stability across changing technologies.
- Distributed data sources can rely on local timestamps at each node, avoiding the need for expensive synchronization in loosely coupled systems.

IV. Frequently Asked Questions

- May data be deleted? Yes, given appropriate policies. Queries may then not be re-executable against the original timestamp anymore (but still against the current timestamp). Landing pages should persist.
- Does the system need to store every query? No. Only data sets that should be persisted for citation / later re-use need to be stored. Persisting queries can be decided individually or policy-based in an automated fashion.
- Can I obtain only the most recent data set? Queries can be re-executed with the original timestamp or with the current timestamp or any other timestamp desired. This allows retrieving the semantically identical data set but incorporating all changes, corrections or updates applied before the given timestamp.
- Which PID system should be used? Any PID system can be applied according to the institutional policy.
- How are the queries created? Queries can either be created manually via an interface/workbench or applications create the proper queries automatically. Both methods require the adaption of the query by adding metadata and timestamps.
- How can I share parts of my database? The query centric view allows selecting any particular view or data subset of the data from the complete data set.
- How does this support giving credit and attribution? Attribution and giving credit is supported via a provenance chain from a subset/view of data to the data set it was derived from, allowing to document intellectual contributions on the way. Analysis and recommendations on how to aggregate bibliometrics and credits is not addressed in the context of this WG.

V. Next Steps

The set of recommendations is undergoing evaluation in a series of pilots in different domains. We encourage interested community members to participate and provide improvements, comments, suggestions and general feedback via the working space of the WG[1]. We are very interested in further real world use cases to act as pilots.

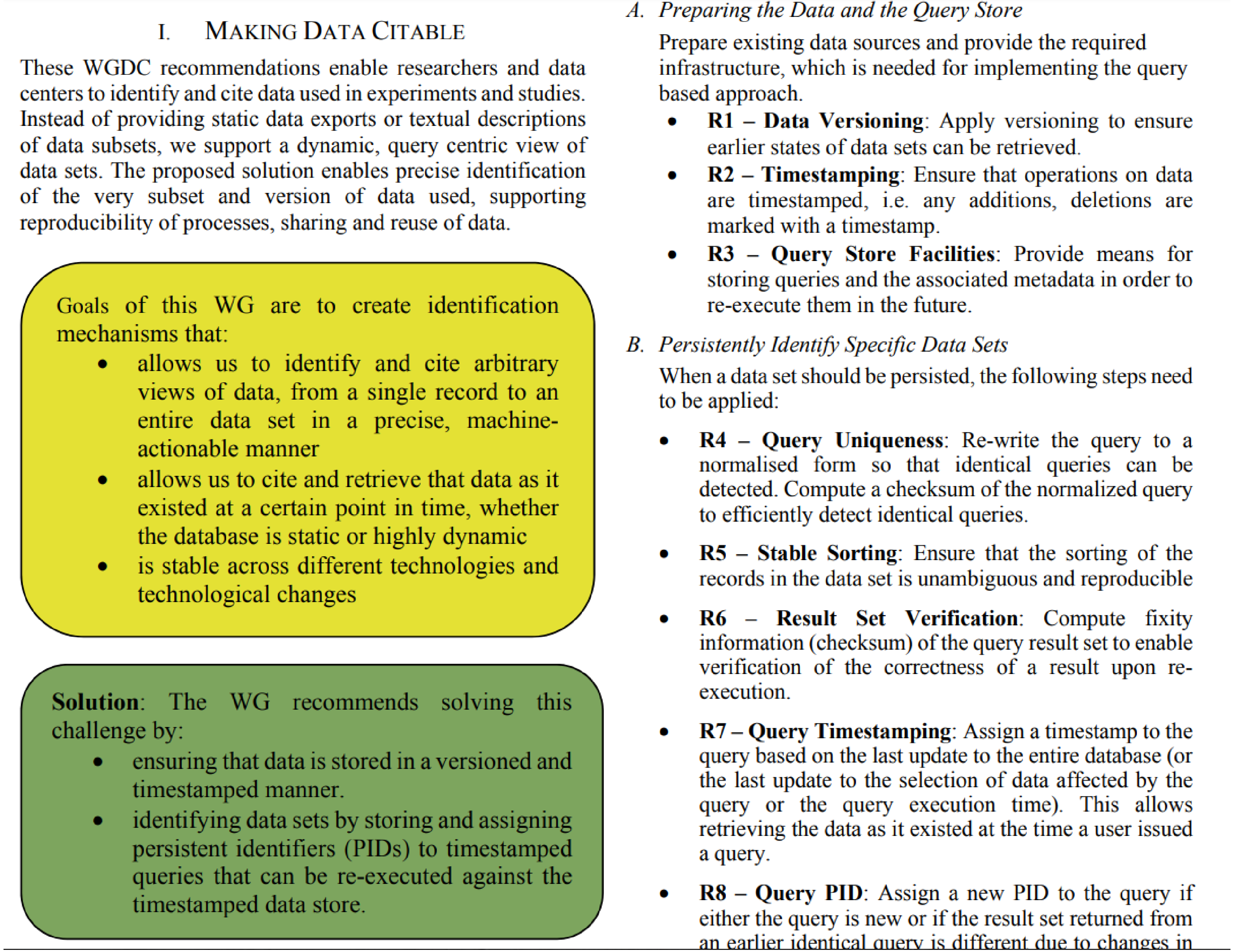
VI. Get Involved

You can find additional information RDA Working Group Page[2]. Please register on the mailing list to stay informed. The community feedback will be collected in the Wiki page[3].

[1] https://rd-alliance.org/group/data-citation-wg/wiki/collaboration-environments.html

[2] www.rd-alliance.org/group/data-citation-wg.html

[3] https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-dynamic-data-citation-recommendations.html



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 subset 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.

A. *Preparing the Data and the Query Store*

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **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 Facilities**: Provide means for storing queries and the associated metadata in order to re-execute them in the future.

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 that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification**: Compute fixity information (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 the time a user issued a query.
- **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

RDA Outputs/Recommendations:
75% or more agree that following Outputs/Recommendations should be our priority (Google sheet order):

1. Scalable Dynamic-data Citation Methodology
2. The FAIRsharing Registry and Recommendations: Interlinking Standards, Databases and Data Policies
3. Metadata Standards Directory
4. 23 Things: Libraries For Research Data

The following Outputs/Recommendations potentially could be analysed (Google sheet order):

1. Basic Vocabulary of Foundational Terminology Query Tool
2. Data Type Model and Registry
3. Workflows for Research Data Publishing: Models and Key Components
4. Legal Interoperability of Research Data: Principles and Implementation Guidelines

When comes to the working groups, the picture is less homogenous. I pulled out following WGs which seem to be in our common interest (Google sheet order):

1. Data Usage Metrics
2. DMP Common Standards
3. FAIR Data Maturity Model
4. FAIRSharing Registry: connecting data policies, standards & databases
5. RDA/WDS Publishing Data Workflows

Following WGs are potentially in our common interest(Google sheet order):

1. Data citation
2. Data versioning
3. Metadata Standards Catalog
4. Research Data Collections
5. Research Data Repository Interoperability

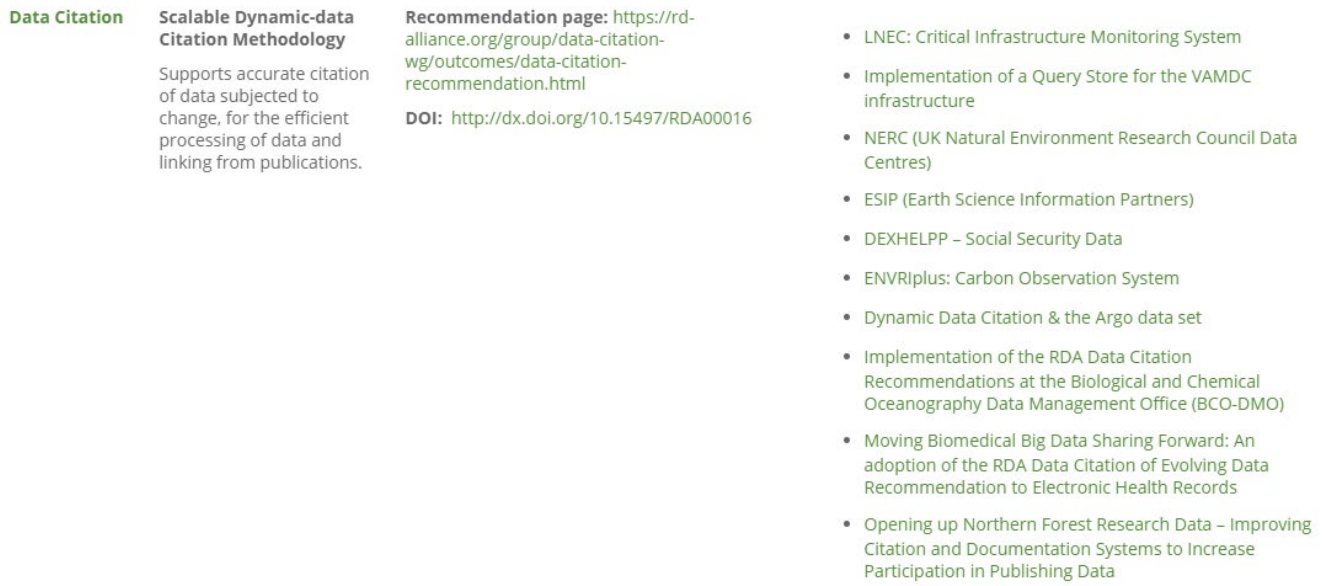
The Interest Groups (IG) survey results are similar to the WG survey.
The following IGs are in our common interest (Google sheet order):

1. Active Data Management Plans
2. Education and Training on handling of research data
3. GO FAIR

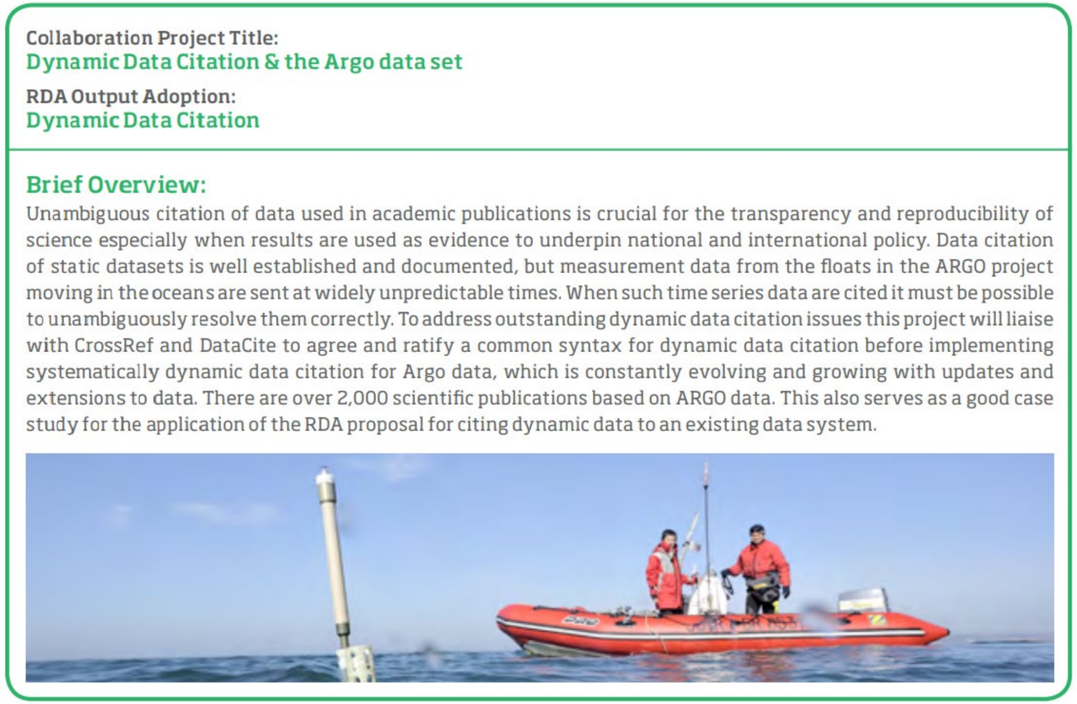
Following IGs are potentially in our common interest (Google sheet order):

1. Archives and Records Professionals for Research Data
2. Data Foundations and Terminology
3. Data policy standardisation and implementation
4. Domain Repositories Interest Group
5. Ethics and Social Aspects of Data
6. Metadata
7. Reproducibility
8. Social Sciences & Humanities Research Data

Adaption stories and use cases



Data Citation	**Scalable Dynamic-data Citation Methodology** Supports accurate citation of data subjected to change, for the efficient processing of data and linking from publications.	**Recommendation page:** https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html **DOI:** http://dx.doi.org/10.15497/RDA00016	• LNEC: Critical Infrastructure Monitoring System • Implementation of a Query Store for the VAMDC infrastructure • NERC (UK Natural Environment Research Council Data Centres) • ESIP (Earth Science Information Partners) • DEXHELPP – Social Security Data • ENVRIplus: Carbon Observation System • Dynamic Data Citation & the Argo data set • Implementation of the RDA Data Citation Recommendations at the Biological and Chemical Oceanography Data Management Office (BCO-DMO) • Moving Biomedical Big Data Sharing Forward: An adoption of the RDA Data Citation of Evolving Data Recommendation to Electronic Health Records • Opening up Northern Forest Research Data – Improving Citation and Documentation Systems to Increase Participation in Publishing Data



Collaboration Project Title:
Dynamic Data Citation & the Argo data set

RDA Output Adoption:
Dynamic Data Citation

Brief Overview:

Unambiguous citation of data used in academic publications is crucial for the transparency and reproducibility of science especially when results are used as evidence to underpin national and international policy. Data citation of static datasets is well established and documented, but measurement data from the floats in the ARGO project moving in the oceans are sent at widely unpredictable times. When such time series data are cited it must be possible to unambiguously resolve them correctly. To address outstanding dynamic data citation issues this project will liaise with CrossRef and DataCite to agree and ratify a common syntax for dynamic data citation before implementing systematically dynamic data citation for Argo data, which is constantly evolving and growing with updates and extensions to data. There are over 2,000 scientific publications based on ARGO data. This also serves as a good case study for the application of the RDA proposal for citing dynamic data to an existing data system.

RDA Magazine 2016

RDA Adoption stories

ADOPTING RDA OUTPUTS FOR CLIMATE DATA MODELLING

DKRZ adopts 6 RDA outputs for climate data modelling

1. **Data Foundation and Terminology**
2. **PID Information Types**
3. **Data Fabric**
4. **Data Type Registries**
5. **Dynamic Data Citation**

Question time ☺

- I might be able to answer …

- Thank you