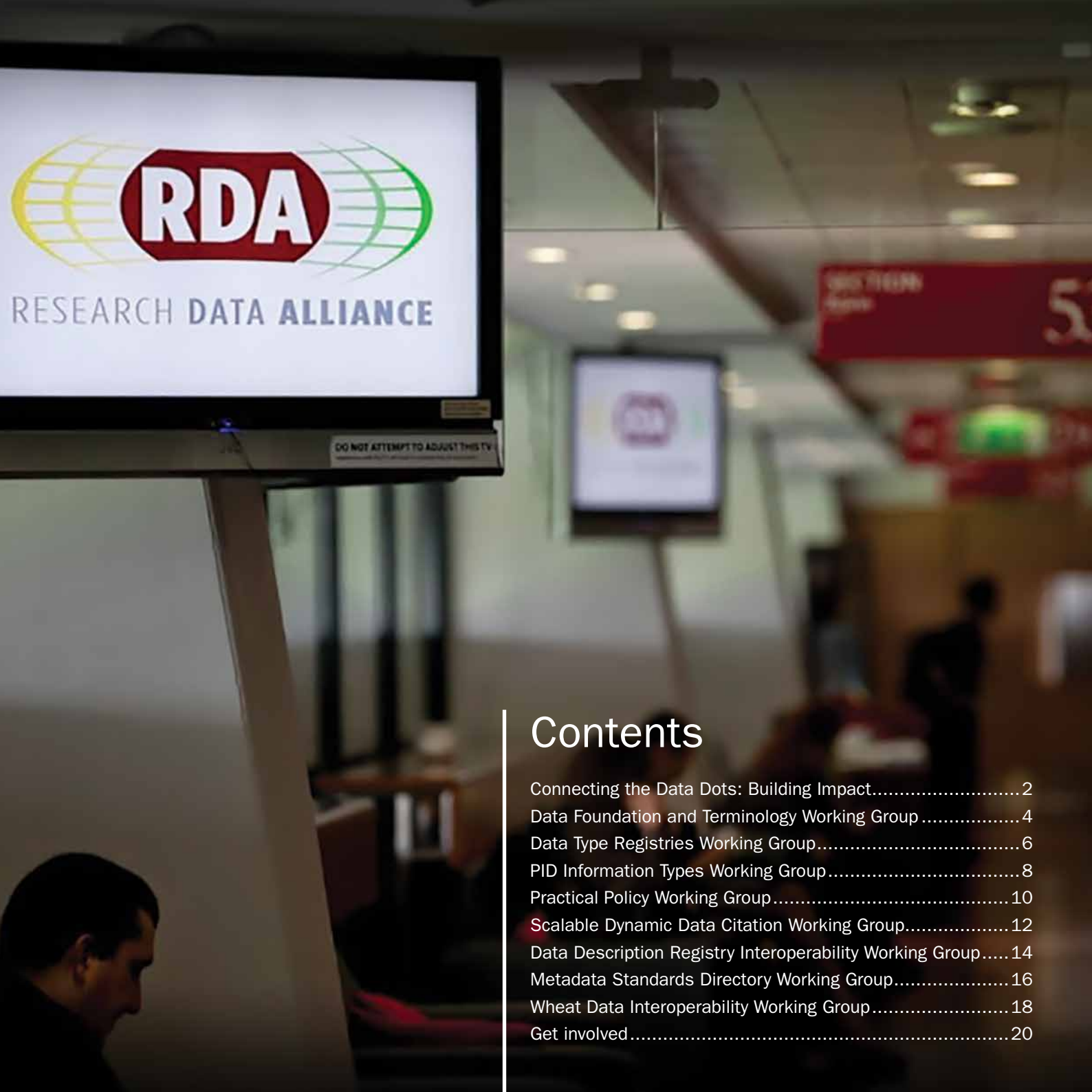
RESEARCH DATA ALLIANCE

OUTPUTS

RESEARCH DATA ALLIANCE



Contents

RESEARCH DATA ALLIANCE

The Research Data Alliance is building the social and technical bridges that enable open sharing of data

"The so called data revolution isn't just about the volume of scientific data; rather, it reflects a fundamental change in the way science is conducted, who does it, who pays for it and who benefits from it. And most importantly, the rising capacity to share all this data – electronically, efficiently, across borders and disciplines – magnifies the impact."

The Data Harvest Report,
John Wood
Chair, Research Data Alliance-Europe,
Co-Chair, RDA Foundation (Global)

Connecting the Data Dots: Building Impact

The Research Data Alliance (RDA)[1] rises to the challenge of changing global data practices by providing concrete solutions to address some of today's many, many data challenges.

Participation in the RDA is open to anyone who agrees to the RDA principles. Data practitioners, community representatives, scientists and technologists come together through focused global Working Groups, exploratory Interest Groups to exchange knowledge, share discoveries, discuss barriers and potential solutions, explore and define policies and test as well as harmonise standards, and recommend pre-existing standards to enhance and facilitate global data sharing. Coupled with this RDA boasts a broad, committed membership of individuals and organizations dedicated to improving data exchange.

Two years since its launch RDA has already published tangible outputs aiming to achieve seamless interoperability, trust, and ultimately to provide growth and employment opportunities by making data re-use less expensive.

So far 8 RDA Working Groups have provided Outputs. Working groups are envisioned as accelerants to data sharing practice and infrastructure in the short-term with the overarching goal of advancing global data-driven discovery and innovation in the long-term.

In the widest sense the group outcomes are pushing forward for:

- New data standards or harmonization of existing standards.
- Greater data sharing, exchange, interoperability, usability and re-usability.
- Greater discoverability of research data sets.
- Better management, stewardship, and preservation of research data.

The 4th RDA Plenary Meeting in Amsterdam (22-24 September 2014) themed "Reaping the fruits" showcased the first concrete outputs from the RDA Working Groups

- **Data Foundation & Terminology**: a model for data in the registered domain.
- **PID Information Types:** a common protocol for providers and users of persistent ID services worldwide.
- **Data Type Registries**: allowing humans and machines to act on unknown, but registered, data types.

  Practical Policy: defining best practices of how to deal with data automatically and in a documented way with computer actionable policy.

The 5th RDA Plenary in San Diego (8-11 March 2015) took important steps forward in facilitating the uptake of the first set of outputs under the "adopt a deliverable" theme as well as marking the launch of the second group of outputs:

- **Metadata standards directory:** Community curated standards catalogue for metadata interoperability
- **Data Citation:** defining mechanisms to reliably cite dynamic data
- **Data Description Registry Interoperability** solutions enabling cross-platform discovery based on existing open protocols and standards
- **Wheat Data Interoperability** impacting the discoverability, reusability and interoperability of wheat data by building a common framework for describing, representing linking and publishing wheat data

In addition the **Data Fabric group** is working with these and other planned outputs to develop a framework for more efficient data management and processing in a loosely coupled manner. This will ultimately aid reproducible data science. All RDA groups are working together to come up with components that will fundamentally change data practices with a wide agreement on turning data into digitally actionable objects, with a persistent identifier and adequate metadata.

Why should these be adopted?

Current data practice challenges are many. Managing, re-using and combining data in science, industry and society is very inefficient, it takes up too much time and binds creative minds. The results produced by data driven work is barely reproducible with an associated lack of trust. A global change of practices is accepted as being an urgent demand, yet there is a severe lack of direction, guidance and trained data experts. Excellent island solutions testing out various options have been developed by different labs, and companies all claiming to have the optimal solutions. Similar to the early Internet this diversity highlights an urgent need for convergence and collaboration.

Adoption of RDA results will lead to:

- Efficient use and re-use of data and reducing related costs
- Increased trust in data science results based on transparent reproducibility.
- Better scientific contribution to society's grand challenges.
- Take up by small companies and entrepreneurs to develop smart data applications for society at large.
- Economic growth & increased employment for data, and other, professionals.

1 www.rd-alliance.org

Who benefits?

Data Citation
Researchers can cite data that is subjected to change. When data gets modified, all changes are reflected in the citation information that includes a time-stamp & version history.

Data Description Registry Interoperability (DDRI)
Infrastructure providers & data librarians to find connections across research data registries and create global views of research data.

Data Foundation & Terminology (DFT)
Scientific Communities through increased cross disciplinary data exchange and interoperability.
Developers by creation of interoperable data management & processing systems.

Data Type Registries (DTR)
Researchers by easily processing or visualising content of unknown data type.
Machines by automatically extracting relevant information from any registered data type.

Metadata
Researchers & service providers to re-use existing standards, to match and map metadata standards leading to interoperability.

PID Information Types (PIT)
Providers by offering a unified access method to all PID service users worldwide.
Developers by supporting just one interface and thus drastically decreasing programming effort.

Practical Policy (PP)
Data managers & scientists by executing documented workflow chains to improve trust.
Researchers by creating reproducible science with the help of documenting procedures.

Wheat Data Interoperability
Data managers & scientists will benefit from the creation of a framework to support the establishment of a global wheat information system.

How do all these dots connect?

Based on similar principles, like those of the Internet community, the Research Data Alliance was started and is run **by practitioners for practitioners** to build social and technical bridges that enable open sharing of data. Through over 60 focused Working Groups (https://www.rd-alliance.org/groups/working-groups) and exploratory Interest Groups (https://www.rd-alliance.org/groups/interest-groups), RDA is working towards making data publishing – the end result of data science - more efficient and developing a complete framework for more efficient data management and processing and ultimately reproducible data science.

Delivering on Promises

RDA's intent is to create deliverables that are developed and used by the community to facilitate data sharing and re-use. Already at this early stage outputs are being adopted by relevant scientific initiatives and organisations in the US and Europe. Through pilot studies they are identifying the potential, limitations and the effort implied in making use of these results for their scientific and infrastructure interests.

Data Foundation and Terminology Working Group

Co-Chairs:
Gary Berg-Cross – Research Data Alliance Advisory Council, Washington D.C.
Raphael Ritz - Max Planck Institute for Plasma Physics
Peter Wittenburg – Max Planck Institute for Psycholinguistics

What is the problem?

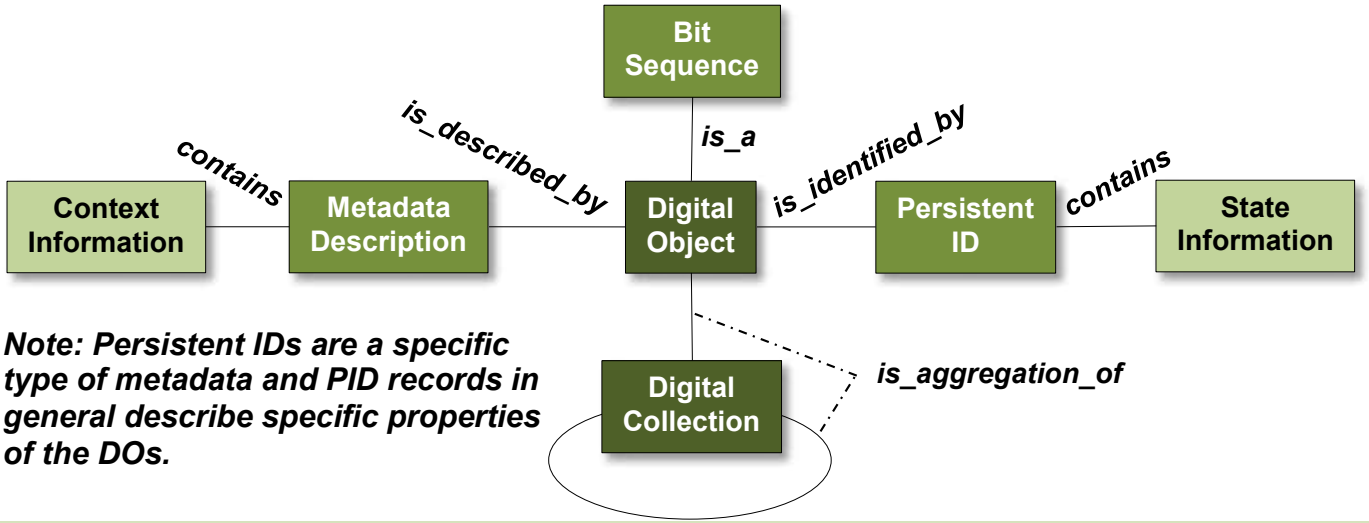
Unlike the domain of computer networks where the TCP/IP and ISO/OSI models serve as a common reference point for everyone, there is no common model for data organisation, which leads to the fragmentation we currently see everywhere in the data domain. Not having a common language between data communities, means that working with data is very inefficient and costly, especially when integrating cross-disciplinary data. As Bob Kahn, one of the Fathers of the Internet, has said, "Before you can harmonise things, you first need to understand what you are talking about."

For the physical layer of data organisations, there is a clear trend towards convergence to simpler interfaces (from file systems to SWIFT-like interfaces[1]). For the virtual layer information, which includes persistent identifiers, metadata of different types including provenance information, rights information, relations between digital objects, etc., there are endless solutions that create enormous hurdles when federating. To give an idea of the scale of the problem, almost every new data project designs yet more new data organisations and management solutions.

We are witnessing increasing awareness of the fact that at a certain level of abstraction, the organisation and management of data is independent of its content. Thus we need to change the way we create and deal with data to increase efficiency and cost-effectiveness.

What are the goals?

- Pushing the discussion in the data community towards an agreed basic core model and some basic principles that will harmonize the data organization solutions.
- Fostering an RDA community culture by agreeing on basic terminology arising from agreed upon reference models.



This diagram describes the essentials of the basic data model that the DFT group worked out in a simplified way. Agreeing on some basic principles and terms would make a lot of difference in data practices.

1 https://wiki.openstack.org/wiki/Swift

When talking about data or designing data systems, we speak different languages and follow different organization principles, which in the end, result in enormous inefficiencies and costs. We urgently need to overcome these barriers to reduce costs when federating data.

What is the solution?

Based on 21 data models presented by experts from different disciplines and about 120 interviews and interactions with different scientists and scientific departments, the DFT WG has defined a number of simple definitions for digital data in a registered[2] domain based on an agreed conceptualisation.

These definitions include:

- **Digital Object** is a sequence of bits that is identified by a persistent identifier and described by metadata.
- **Persistent Identifier** is a long-lasting string that uniquely identifies a Digital Object and that can be persistently resolved to meaningful state information about the identified digital object (such as checksum, multiple access paths, references to contextual information etc.).
- **A Metadata description** contains contextual and provenance information about a Digital Object that is important to find, access and interpret it.
- **A Digital Collection** is an aggregation of digital objects that is identified by a persistent identifier and described by metadata. A Digital Collection is a (complex) Digital Object.

A number of such basic terms have been defined and put into relation with each other in a way that can be seen as spanning a reference model of the core of the data organisations.

What is the impact?

The following benefits will come from wide adoption of a harmonized terminology:

- Members of the data community from different disciplines will be able to interact more easily with each other and come to a common understanding more rapidly.
- Developers can design data management and processing software systems enabling much easier exchange and integration of data from their colleagues in particular in a cross-disciplinary setting (full data replication for example could be efficiently done if there is an agreement on basic organization principles for data).
- It will be easier to specify simple and standard APIs to request useful and relevant information related to a specific Digital Object. Software developers would be motivated to integrate APIs from the beginning and thus facilitate data re-use, which currently is almost impossible without using information that is exchanged between people.
- It will bring it a step closer to automating data processing where all can rely on self-documenting data manipulation processes and thus on reproducible data science.

When can this be used?

The definitions have been discussed at RDA 4th Plenary meeting (September 2014) and are available as a document and on a semantic wiki to invite comments and usage since January 2015. RDA and the group members will take care of proper maintenance of the definitions. For more information see

https://rd-alliance.org/group/data-foundation-and-terminology-wg.html

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

In the next phase of the work, more terms will be defined and interested individuals will have the opportunity to comment via the semantic wiki.

2 There will always exist data in private, temporary stores, which will not be made accessible in a standard way.

Data Type Registries Working Group

Co-Chairs:
Larry Lannom - Corporation for National Research Initiatives,
Daan Broeder - Max Planck Institute for Psycholinguistics

What is the problem?

Often researchers receive files from colleagues, follow links, or otherwise encounter data created elsewhere that they would like to make use of in their own work. However, they may not know how to work with it, interpret it or visualise its content, if they are unfamiliar with the specifics of the structure and/or meaning of the data. Frequently, researchers end up not using such data, since it requires extra work to look for explanations and tools, (and install these tools where necessary) – so that they can access the data.

What are the goals?

The aim of the Data Type Registries Working Group (DTR WG) was to allow data producers to record the implicit details of their data in the form of Data Types and to associate those Types, each uniquely identified, with different instances of datasets.

Linking data type identifiers to datasets will provide, data consumers with an indication of the type of datasets they encounter. This means being able to determine which services (and other useful information) to use, to understand and to process the data, without additional support from the respective data producers. DTRs are meant to provide machine-readable information, in addition to presenting human readable information.

What is the solution?

DTRs offer developers or researchers the ability to add their type definitions in an open registry and, where useful, add references to tools that can operate on them. For example, a user who received an unknown file could query a DTR and receive back a pointer to a visualisation service able to display the data in a useful form.

A fully automated system could use a DTR, much like the MIME type system enables the automatic start of a video player in the browser once a video file has been identified. We envision humans taking advantage of Data Types in DTRs through the type definitions that clarify the nuanced and contextual aspects of structured datasets.

Precise typing of data sets and collections, combined with one or more registries that define those types in a standard fashion, would benefit every sector of data management, especially interoperability and reuse.

Data Types in DTRs can be used to extend or expand existing types, e.g., MIME types, which provide only container-level parsing information. They can additionally describe experimental context, relationships between different portions of data, and so on. Data Types are deliberately intended to be quite open in terms of registration policies.

The DTR solution is particularly useful for:

- Researchers dealing with data in a cross-disciplinary, cross-border context, who encounter unknown data types. Using the DTR service allows them to immediately process and/or visualize the content of such data types"
- Machines that want to extract the checksum information of a data object from a PID record to check whether the content is still the same. Without knowing the details of the PID service provider, the machine could ask for checksum for example, since this is an information type which all PID service providers agreed upon and registered in the DTR.

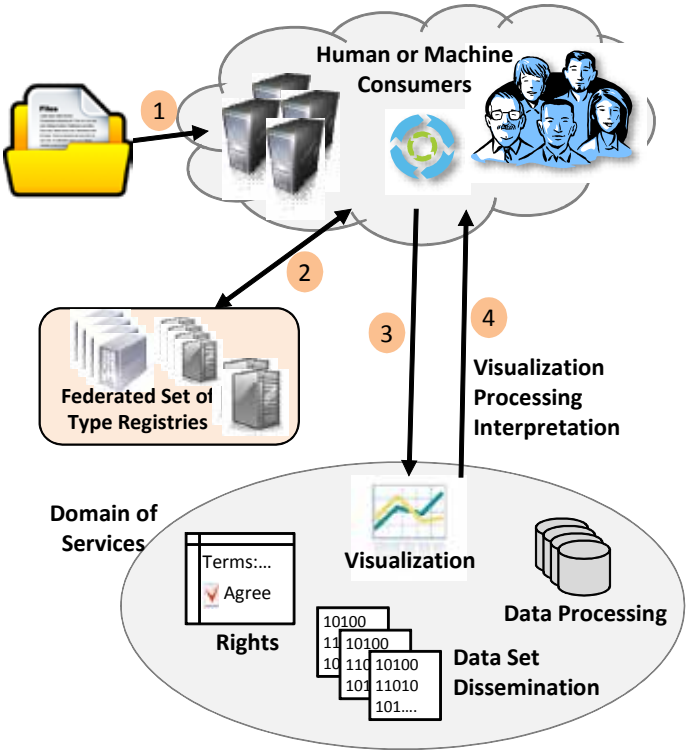
What is the impact?

The potential impact on scientific practices is substantial. Unknown data types as described above can be exploited without any prior knowledge and thus an enormous gain in time and/or in interoperability can be achieved. In a similar way to the MIME types that allow browsers to automatically select visualization software plug-ins when confronted with a certain file type extension, scientific software can make use of the definitions and pointers stored in the DTR to continue processing without the user acquiring knowledge beforehand.

DTRs pave the way to automatic processing in the data domain, which is becoming increasing complex, without putting an additional load on the researchers.

However, the individuals who categorize data types, are required to enter the associated, relevant information into a DTR.

It is assumed that there will be a federation of such DTRs setup to satisfy different needs.



This diagram illustrates how the Data Type Registry (DTR) works. A user or machine receives an unknown type (1) which can be a file or a term, for example. The DTR is contacted and returns information about an available service (2) this allows the user or machine to continue processing the content (3, 4) such as visualizing an image without asking prior knowledge from the user. This makes cross-disciplinary and cross-border work much more efficient and enables data driven science even to those who are not data experts.

When can this be used?

The first groups are building software to implement such a DTR concept and make the software available. The RDA PID Information Type (PIT) Working Group is already using the first DTR prototype version in its API. The latest version of a DTR prototype is available here: http://typeregistry.org/. Please check the information on the DTR WG's web page at https://www.rd-alliance.org/group/data-type-registries-wg.html for updates.

This simple model will be the start for designing DTRs, with the intention to extend the specifications according to priorities and usage.

PID Information Types Working Group

Co-Chairs:
Tobias Weigel – DKRZ
Timothy DiLauro – John Hopkins University

What is the problem?

Numerous systems and providers to register and resolve Persistent Identifiers (PIDs) for Digital Objects and other entities have been designed in the past and are used today. However, almost all of them differ in the way they allow researchers to associate additional information, such as for proving identity and integrity with the PID. For application developers this is an unacceptable situation, since for all providers a different Application Programming Interface (API) needs to be developed and maintained. If a researcher finds a useful file and wishes to check that it is still the same stream of bits, as when it was first created, the researcher should be able to request the checksum independent of the provider holding the PID. How should the researcher do this not knowing whether the provider offers this information and if so, how to request it? We can overcome such extreme inefficiencies only if all providers agree on a common API, register their information types in a common data type registry and agree on some core types, such as the checksum.

What are the goals?

The aim of the PID Information Types Working Group (PIT) was to :

- Come to a core set of information types and register (and define) them in a commonly accessible Data Type Registry
- Provide a common API and prototypical implementation to access PID records that employ registered types

What is the solution?

The PIT Working Group accomplished the following:

- Defined and registered a number of core PID information types (such as checksum)
- Developed a model to structure these information types
- Provided an API, including a prototypical server implementation that offers services to request certain types associated with PID records by making use of registered types.

Due to high demand, a variety of trusted PID service providers have been set up already, yet all of the different attributes associated with the registered PIDs make the life of a software developer a nightmare. It is essential to harmonize the major information types and suggest a common API, so that if the checksum is requested one has to program one piece of software independent of the provider.

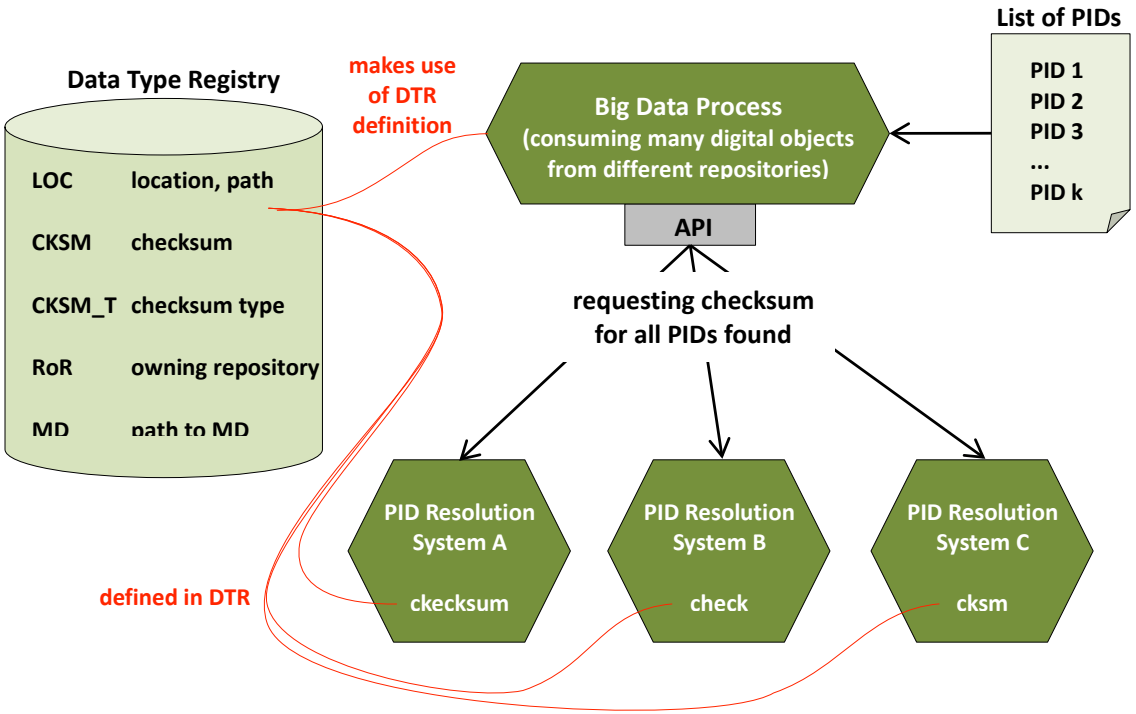
The set of core information types currently provided can help to illustrate cross-discipline usage scenarios. It can also act as an example for a community-driven governance process creating and governing more user-driven types. PID service providers and community experts need to come together regularly and add types to the data type registry to make full use of the possibilities of the results of the PIT group.

It is now essential to convince PID service providers such as those using the Handle System (DOI, EPIC, etc.) to adopt the API to unify access. The diagram gives an example of the usage and potential of the suggested solution.

What is the impact?

It is important to envisage the situation in a few years, when the amount and complexity of data has been increased in all sciences and there is a greater need to rely on automatic processes, as human intervention means loss of efficiency. In such scenarios, particularly in the area of big data analytics, communities can exploit the wealth of the data domain by relying on semantic interoperability between all relevant actors. The above example is just one small usage scenario that would be enabled if the relevant PID service providers accept the results of the PIT WG and harmonize their approach. Application software writing would be reduced dramatically since only one API would be supported and one module would be sufficient for retrieving the checksum, for example, and checking identity and integrity.

The strengthening of PID information types could also move the existing identifier systems and the overall idea of identification into a more central and fundamental position as suggested by DFT's core model of a Digital Object, leading to an enormous increase in efficiency when dealing with data.



Assume that you have a list of PIDs referring to data that you would like to use in a computation. Despite the fact that the PIDs might be registered at various providers, you would simply use a single module that reads (or 'selects') the relevant PID from the list of PIDs, and then submits a request to the appropriate resolver to send the checksum.

If all actors refer to the same entry in the DTR, interoperability is a given. That is, one module would be sufficient to retrieve the checksums, independent of the internal terminologies used by the various providers.

When can this be used?

Initial work has already been done on building software to implement a first prototype based on the defined PIT API. This first prototype works together with the DTR prototype and both are publicly available, but not designed for production use.

Please check the information and updates on the PIT group's web page at https://www.rd-alliance.org/group/pid-information-types-wg.html.

It is now time to convince the PID service providers to adopt the solution.

Practical Policy Working Group

Co-Chairs:
Reagan Moore, RENCI
Rainer Stotzka, Karlsruhe Institute of Technology

What is the problem?

Repositories' responsibilities for data stewardship and processing require a highly automated, safe and documented management strategy. Management policies need to be enforced, administrative policies need to be automated, and assessment validation policies need to be evaluated periodically.

With the increasing amount and complexity of data, repositories need to publish their policies and procedures to build trust in their operation. By sharing policies, repositories can build upon discipline expertise, and implement improved procedures for ensuring trustworthiness.

Operations or chains of operations that are computer actionable and enforced on collections of data objects can be based on the outcomes from the "Practical Policy" (PP) working group. The outcomes are stated in natural languages and can be turned into robust and tested executable procedures. The ability to re-execute procedures is at the basis of reproducible science, an important element in the chain of building trust and one of the core elements in repository certification processes.

What are the goals?

The goals of the PP Working group were to:

- Define computer actionable PPs that enforce proper management and stewardship, automate administrative tasks, validate assessment criteria, and automate types of scientific data processing
- Identify typical application scenarios for practical policies such as replication, preservation, metadata extraction, etc.
- Collect, register and compare existing practical policies
- Enable sharing, revising, adapting and re-use of such practical policies and thus harmonize practices, learning from good examples and increasing trust

Since these goals were broad in scope, the PP WG focused its efforts on a few application scenarios for the collection and registration process.

Current practice in managing and processing data collections are determined by manual operations and ad-hoc scripts making verification of the results an almost impossible task. Establishing trust and a reproducible data science requires automatic procedures which are guided by practical policies. Collecting typical policies, evaluating them and providing best practice solutions will help all repositories and researchers.

What is the solution?

In order to identify the most relevant areas of practice, the PP WG conducted a survey as a first step. The analysis of the survey resulted in 11 highly important policy areas which were tackled first by the WG: 1) contextual metadata extraction, 2) data access control, 3) data backup, 4) data formal control, 5) data retention, 6) disposition, 7) integrity (incl. replication), 8) notification, 9) restricted searching, 10) storage cost reports, and 11) use agreements.

Participants and interested experts were asked to describe their policy suggestions in simple semi-formal descriptions. With this information, the WG developed a 50-page document covering the simple descriptions, the beginning of a conceptual analysis and a list of typical cases such as extract metadata from DICOM, FITS, netCDF or HDF files.

The WG functioned through RDA 5th Plenary (March 2015), and focused on further analysing, categorising and describing the offered policies. Volunteers reviewed the policies and different groups implemented some of these policies in environments such as iRODS and GPFS. The goal was to register prototypical policies with suitable metadata so that people can easily find what they are looking for and re-use what they found at abstract, declarative or even at code level. At this point, there is still much work to be done to reach a stage where the policies can be easily re-used. An initial template has been developed that describes the constraints that control the policy, the state information needed to evaluate the constraints, the operations that are performed by the policy, and the state information needed to execute the operations.

What is the impact?

The potential impact is huge. In the ideal case, data managers or data scientists can simply plug-in useful code into their workflow chains to carry out operations at a qualitatively high level. This will improve the

quality of all operations on data collections and thus increase trust and simplify quality assessments. Large data federation initiatives such as EUDAT(http://eudat.eu) and the DATANET Federation Consortium (US) (http://datafed.org)are very active in this group, since they also expect to share code development and maintenance, thus saving considerable effort by re-using tested software components. Research Infrastructure experts that need to maintain community repositories can simply re-use best practice suggestions, thus avoiding ending up in traps. In particular, when these best practice suggestions for practical policies are combined with proper data organisations, as suggested by the Data Foundation and Terminology Working Group, powerful mechanisms will be in place to simplify the data landscape and make federating data much more cost-effective.

When can this be used?

The document mentioned above already provides a valuable resource to get inspiration and perhaps make use of suggested policies, therefore improving people's own ideas or to even profit from developed code.

Once evaluated, properly categorised and described, the next step ahead will be registering practical policies in suitable registries, so that data professionals can easily re-use them, if possible even at code level. The group intends to progress to this step for a number of policy areas, making use of the policy registry developed by EUDAT.

Policies are expected to form an essential component of the Data Fabric Interest Group outcomes. Federation of existing data repositories depends upon the ability to characterize assertions about each participating collection, and enforce the assertions across the participating repositories.

Example assertions include:

- Presence of required descriptive metadata
- Presence of required derived data products (typically alternate data formats)
- Guarantees on integrity
- Guarantees on data provenance
- Logical arrangements that span repositories (virtual collections)
- Guarantees on access controls.

Policies provide a way to quantify the management steps needed to enforce an assertion, share the management step with other repositories, and automate enforcement. The Data Fabric Interest Group can promote the policies needed to manage repository federations.

Within the DataNet Federation Consortium, a "Policy Workbook" is being created that extends the policy set defined in the PP Working Group. The "Policy Workbook" will be published through the iRODS Consortium.

For more details on the PP WG, see https://www.rd-alliance.org/group/practical-policy-wg.html

The diagram indicates the final goal of the PP WG. A policy inventory will be made available with best practice examples. Data managers will have the ability to select and implement the procedures most relevant to them.

Scalable Dynamic Data Citation Working Group

Co-Chairs:
Andreas Rauber, Vienna University of Technology
Dieter Van Uytvanck, CLARIN
Ari Asmi, University of Helsinki
Stefan Pröll, SBA Research (Secretary)

What is the problem?

Digitally driven research is dependent on quickly evolving technology. As a result, many existing tools and collections of data were not developed with a focus on long term sustainability. Researchers strive for fast results and promotion of those results, but without a consistent and long term record of the validation of their data, evaluation and verification of research experiments and business processes is not possible.

To verify research results, repeat studies, or perform meta-studies reusing data, the data used needs to be precisely identified. This, however, is complicated by two challenges: (1) Especially in big data settings, researchers rarely use an entire dataset. Instead, they select specific subsets /views of the entire dataset based on their individual requirements, such as a specific time-range, a set of measurements, etc. (2) Data is not static: new data are often added to datasets, and erroneous values are often corrected or deleted from datasets. This makes it difficult to identify precisely which data (or which version of the dataset) was cited, over time. Thus, there is a strong need for data identification and citation mechanisms that identify arbitrary subsets of large data sets with precision in a machine-actionable way. These mechanisms need to be user-friendly, transparent, machine-actionable, scalable and applicable to various static and dynamic data types.

What are the goals?

The aim of the Dynamic Data Citation Working Group was to devise a simple, scalable mechanism that allows the precise, machine-actionable identification of arbitrary sub selections of data at a given point in time irrespective of any subsequent addition, deletion or modification. The principles must be applicable regardless of the underlying database management system (DMBS), working across technological changes. It shall enable efficient resolution of the identified data, allowing it to be used in both human-readable citations as well as machine-processable linking to data as part of analysis processes.

What is the solution?

The approach recommended by the Working Group relies on dynamic resolution of a data citation via a time-stamped query also known as dynamic data citation. It is based on time-stamped and versioned source data and time-stamped queries utilized for retrieving the desired dataset at the specific time in the appropriate version.

The solution comprises of the following core recommendations:

- Data Versioning: For retrieving earlier states of datasets the data needs to be versioned. Markers shall indicate inserts, updates and deletes of data in the database.
- Data Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- Data Identification: The data used shall be identified via a PID pointing to a time-stamped query, resolving to a landing page.

Although the exact technical implementation depends on existing

local data structures and procedures, evaluations of numerous pilot projects involving various data types (SQL, CSV, XML) indicate the applicability and versatility of this solution.

The WG recently created the RDA recommendations for data citation, which is available as a draft on the RDA Website. The document provides 13 recommendations providing guidance from preparing the data store via the persistent identification of datasets, the retrieval of a dataset until the long term perspective for identifiable datasets.

What is the impact?

The main impact of this solution is to provide a mechanism supporting reproducibility of scientific research by allowing for a data source to be dynamically updated when information is added, updated or deleted, while still enabling for the reproduction of any previous or intermediate version of the data. The approach detailed above has several advantages over current practices, which mainly utilize redundant data deposits or ambiguous natural language textual descriptions.

First, the query/expression identifying the dataset provides valuable provenance information on the way the specific dataset was constructed, as opposed to merely having a data dump.

Secondly, the recommended solution allows users to re-execute the query with the original time stamp and retrieve the original data, or to obtain the current version of the data with all additions and corrections by executing it against the current version of the data repository. This allows them to compare the resulting differences.

Thirdly, it is generally 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).

As data migrates to new representations, the queries can also be migrated, ensuring stability across changing technologies.

By promoting a consistent approach, decision making and scientific research based on data will become more transparent and reproducible.

When can this be used?

As demonstrated by first successful pilots, this approach can be applied right now. The recommendations are available for comments and can be used as an implementation guideline.

For more information on the solutions detailed above or to learn more about the Dynamic Data Citations Working Group, please visit https://rd-alliance.org/groups/data-citation-wg.html.

Data Description Registry Interoperability Working Group

Co-Chairs:
Amir Aryani, Australian National Data Service
Adrian Burton, Australian National Data Service

What is the problem?

In recent years there has been a significant growth of research data repositories and registries; however, these infrastructures are fragmented across institutions, countries and research domains. As such, finding research datasets is not a trivial task for many researchers.

What are the goals?

Data Description Registry Interoperability WG is working on a series of bi-lateral information exchange projects and an open, extensible, and flexible cross-platform research data discovery software solutions.

Where research data registries and repositories provide machine-to-machine readable interfaces, the issue of wider discovery is often addressed either by metadata aggregation or federated search. However, the main problem is providing scientists search results for datasets that are actually relevant to their research. Such relevance depends on research context, and as a result enabling cross-platform discovery includes providing a connected graph of researchers, research activities (projects and grants), research datasets, publications and other research outcomes and research concepts.

This working group does not aim for a monolithic solution, avoiding a one uber-portal to rule them all. Rather it compiles simple enabling infrastructures based on existing open protocols and standards with a flexible and extensible approach that allows registries to opt-in and enables any third-party to create particular global views of research data.

Who is involved in this working group?

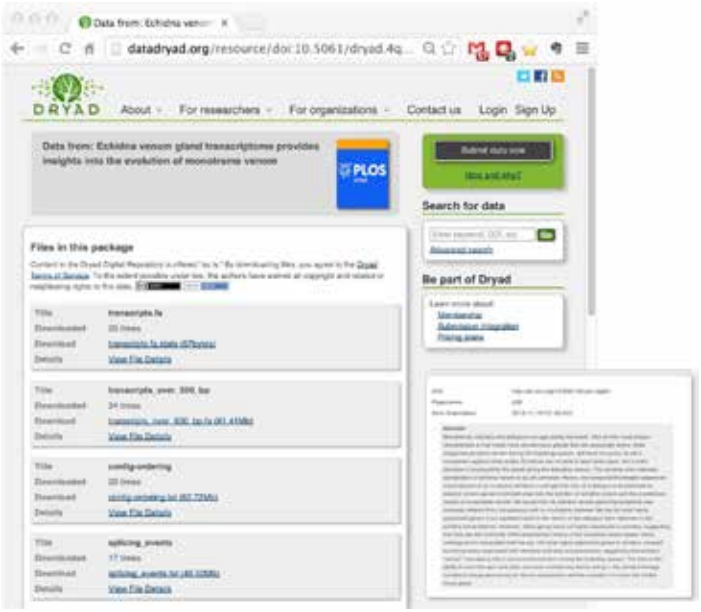
The outcome and the deliverables of this working group will be the result of the direct contribution of the following major institutions in Australia, US and Europe: Australian National Data Service (ANDS), CERN, DANS, DataCite, DataPASS, da-ra, Dryad, Thomson Reuters DCI, VIVO Cornell.

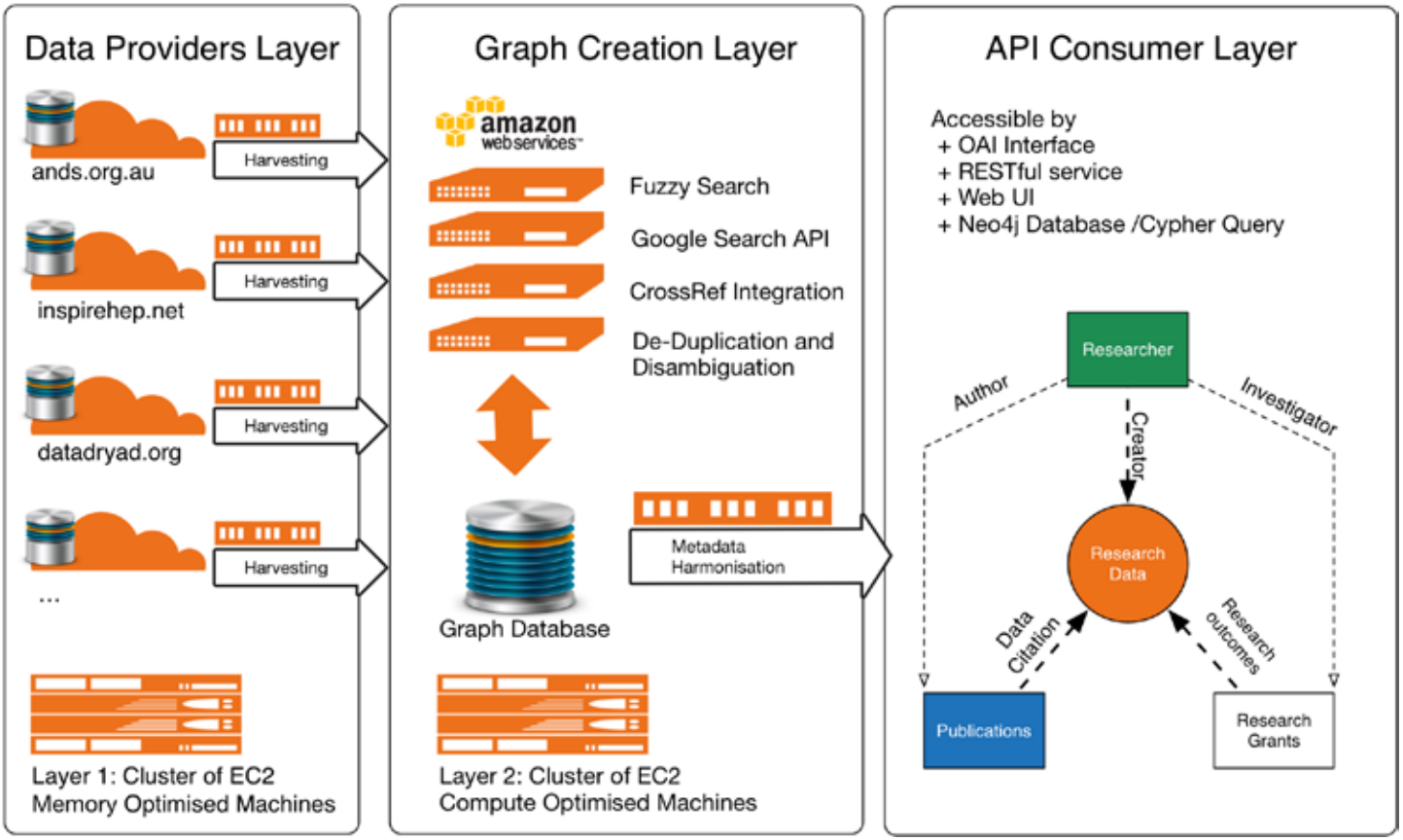
What is RD-Switchboard?

Research Data Switchboard is a collaborative project by the members of the DDRI WG. This project leverages DataCite DOI, ORCID and other persistent identifiers, and uses simple but effective research graph technology to link datasets. This system currently links datasets across the following platforms: Dryad, INSPIREHEP (at CERN), ORCID, Figshare and link Australian research datasets through Research Data Australia - supported by ANDS.

For example, this platform enables connecting this dataset by Associate Professor Katherine Belov: Wong ESW, Nichol S, Warren WC, Belov K (2013) Data from: Echidna venom gland transcriptome provides insights into the evolution of monotreme venom. Dryad Digital Repository http://dx.doi.org/10.5061/dryad.4qq0v to her other data collections in Research Data Australia:

- Tammar wallaby thymus transcriptomes http://researchdata.ands.org.au/tammar-wallaby-thymus-transcriptomes-dataset/11126
- IDMM Immunome Database for Marsupials and Monotremes http://researchdata.ands.org.au/idmm-immunome-database-for-marsupials-and-monotremes/11139.





The figure above shows the functions of the three layers of Research Data Switchboard:

Provider Layer: This layer enables data providers to import metadata records into the platform using OAI-PMH or RESTful services.

Graph Creation Layer: This layer aggregates information, and uses Google API and other services to identify missing connections.

API Consumer Layer: This layer enables e-Infrastructure providers and university librarians to find connections across research data registries.

When can be this be used?

The work on Research Data Switchboard will continue in the scope of the Data Description Registry Interoperability Working Group. The upcoming RDA Plenaries will provide momentum and opportunity for new partners to join and work toward a sustainable and innovative interoperability platform.

Metadata Standards Directory Working Group

Co-Chairs:
Alex Ball, UKOLN Informatics
Jane Greenberg, Metadata Research Center
Keith Jeffery, Keith G Jeffery Consultants
Rebecca Koskela, DataONE

What is the problem?

When working with research datasets, a common challenge is the information within them is often difficult to identify, contextualize, interpret and use due to the inconsistent approaches in applying related metadata, or metadata schemes. To fully understand the content within datasets, researchers need metadata that clearly describes, explains, and associates the dataset with various other entities.

However, metadata needs vary depending on the data type and the application. This results in the use of numerous metadata schemes and lack of interoperability[1]. With the continued use of custom metadata schemes, and the development of rival, incompatible standards, there are now even more barriers to interoperation[2].

This challenge can be overcome through the implementation of one set of metadata standards, which would involve the application of the same metadata, and hence data, in multiple contexts and systems.

A collaborative, open directory of metadata standards applicable to scientific data can help address these infrastructural challenges, by allowing researchers to:

- Learn about the various metadata standards applicable to their research;
- Learn about controlled vocabularies used by their community;Understand the elements that comprise these standards and vocabularies; and
- Map between elements when combining data from different sources.

These standards can only be successful if they are user-friendly, well promoted and widely adopted in target communities.

What are the goals?

The goals of this group are three-fold:

1. Set up a sustainable, community-driven RDA Metadata Standards Directory, designed for users rather than automated tools, which provides brief details for common research data.

2. Compile a set of use cases that analyze and document the various ways in which metadata can be used (e.g. for discovery, exchange, re-use, etc.).

3. Lay the foundation for a future RDA Working Group to develop a machine-understandable catalogue of metadata standards.

What is the solution?

The United Kingdom Digital Curation Centre (DCC) launched a Disciplinary Metadata Standards Catalogue (http://www.dcc.ac.uk/resources/metadata-standards) just before this Working Group started its activity. The DCC's catalogue was adopted, enriched, and expanded by the Working Group.

The Working Group developed a functional prototype directory (http://rd-alliance.github.io/metadata-directory/), based around the GitHub infrastructure, that places the information from the DCC directory into an environment where it can be maintained transparently and with full version control.

Metadata use cases were also collected from Working Group members using a standard template and ultimately included in the set of use cases compiled by the RDA Metadata Interest Group.

1 Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. PLoS ONE, 6(6), e21101. doi:10.1371/journal.pone.0021101

2 Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505–1520. doi:10.1002/asi.22683

What is the impact?

The RDA Metadata Standards Directory has many benefits for the community:

- By guiding researchers towards the metadata standards and tools relevant to their discipline, the directory drives up adoption of those standards, improving the chances of future researchers finding, accessing, and reusing the associated data.
- By raising awareness of existing standards, the directory reduces the proliferation of ***ad hoc*** metadata formats and helps direct future standards development efforts towards those areas that most need it.
- If a topical standard is not available, the directory allows researchers to look beyond their subject boundaries for standards that are a close fit for their work.
- By raising awareness of standards among tool developers, the directory can help improve technical support for those standards

The human-readable directory is also the first step towards a machine-understandable catalogue, which would have a significant impact on the ability of researchers and service providers to migrate metadata automatically between systems. Through this automation, services would be allowed to bring together specific data based on smart metadata selection, thereby breaking down barriers in research and opening up new possibilities for startup companies and entrepreneurs.

When can this be used?

The DCC directory has been available for use since May 2012. RDA's prototype directory is fully functional, open to the community, and actively monitored so that contributions are fed back to the DCC version and vice versa.

For more information on the usage of this metadata standards directory, please consult the online documentation (http://rd-alliance.github.io/metadata-directory/) on GitHub or a recent article on this work[3].

3 Ball, A., Chen, S., Greenberg, J., Perez, C., Jeffery, K., & Koskela, R. (2014). Building a Disciplinary Metadata Standards Directory. International Journal of Digital Curation 9(1), 142–151. doi:10.2218/ijdc.v9i1.308

Wheat Data Interoperability Working Group

Co-Chairs:
Esther Dzale Yeumo Kabore, INRA
Richard Fulss, CIMMYT

Interoperability of all wheat-related data

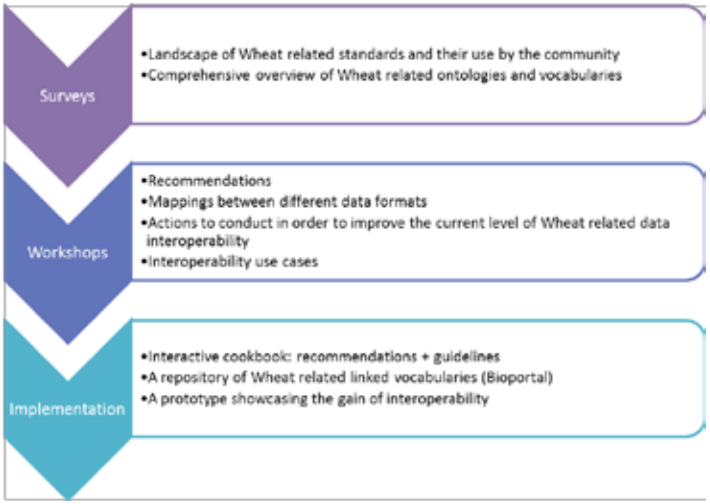
What is the problem?

The Wheat Data Interoperability Working Group (WDIWG) is working within the global context of a large societal challenge, due in part to the following:

- Wheat is the most widely grown crop in the world
- Wheat provides 20% of the world's daily protein and calories
- Wheat is the second most important crop in the developing world after rice
- Wheat production has not satisfied demand in recent years
- It is expected that by 2050 the demand for wheat will increase by 60%

To respond to these facts – and to produce an adequate amount of wheat – the yield increase must go from 1% a year to 1.6% a year.

In order to tackle this issue, many organizations and initiatives are doing research in experimental and farmers' fields, as well as in laboratories, ultimately generating a large quantity of heterogeneous data that are stored in different systems/platforms/repositories. The WDIWG considers data standards harmonization a priority in promoting interoperable wheat data.



What are the goals?

The goals of the WDIWG are to make wheat data interoperable by agreeing on a common set of:

- Metadata standards
- Data formats
- Vocabularies
- Guidelines for describing, representing, and linking data

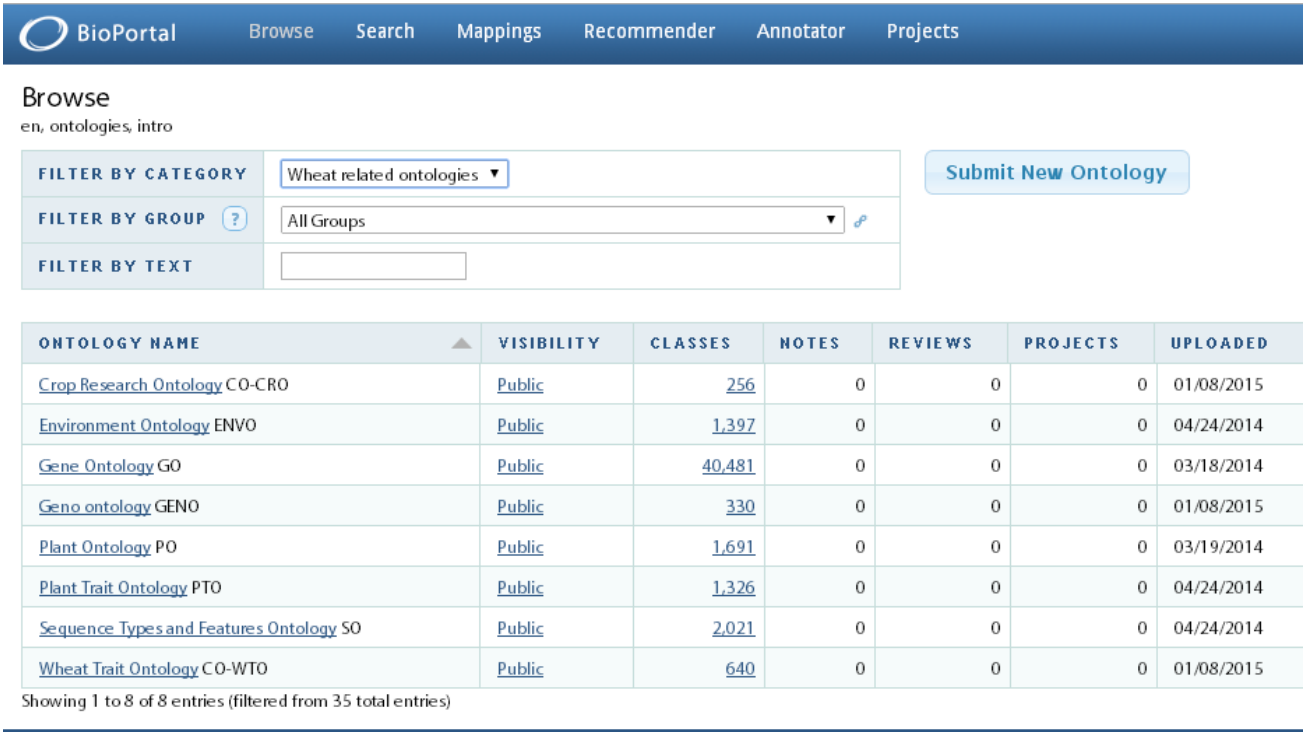
Furthermore, the group aims to produce tools that encourage the adoption of the recommendations and guidelines.

Note that the group did not start from zero, the community has a large amount of assets which are used as a basis. The requirements for the work are based on the real needs of the wheat community.

What is the solution?

The needs of the wheat community are addressed in three ways:

- By building an interactive cookbook with recommendations and guidelines on data formats and standards to use,
- By identifying wheat-related vocabularies and ontologies and including them in a single human and machine readable portal,



» By building a prototype based on real use cases that leverage the recommendations in order to assess the gain of interoperability.

What is the impact?

The impact of this work is the immediate and ongoing improvement of discovery, reusability, and interoperability of data within the wheat community.

Going forward, the standardization and harmonization of wheat data will reduce variability and increase the relevance of wheat data related tools.

The outputs of this group have been adopted by the WheatIS (www.wheatis.org) which is an effort to build an international Wheat Information System.

When can this be used?

The guidelines produced by the group, as well as the bioportal of wheat-related linked vocabularies, are directly usable now.

Following the guidelines and linking into existing vocabularies will give wheat related data a larger relevance and impact going forward.

For more information on WDIWG visit https://www.rd-alliance.org/groups/wheat-data-interoperability-wg.html.

See also http://ist.blogs.inra.fr/wdi/recommendations-for-phenotypes/ for direct links to clear recommendations.

Get involved

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

RDA Mission: The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data.

RDA Guiding Principles

- **Openness** – Membership is open to all interested individuals who subscribe to the RDA's Guiding Principles. 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.

How to play a part in the RDA Process

There are severals way in which you can play a part in RDA:

- **Register** to the website and become a member of the RDA community - https://www.rd-alliance.org/user/register
- **Join the discussion** and subscribe to the RDA Working and Interest Groups - https://www.rd-alliance.org/groups
- **Participate** in the next Bi-annual Plenary Meeting - https://rd-alliance.org/plenary-meetings/rda-sixth-plenary-meeting.html

https://www.rd-alliance.org

Contact: enquiries@rd-alliance.org

Photography by Inge Angevaare, Johnny Babmbury
Designed and produced by RDA Europe (May 2015).