



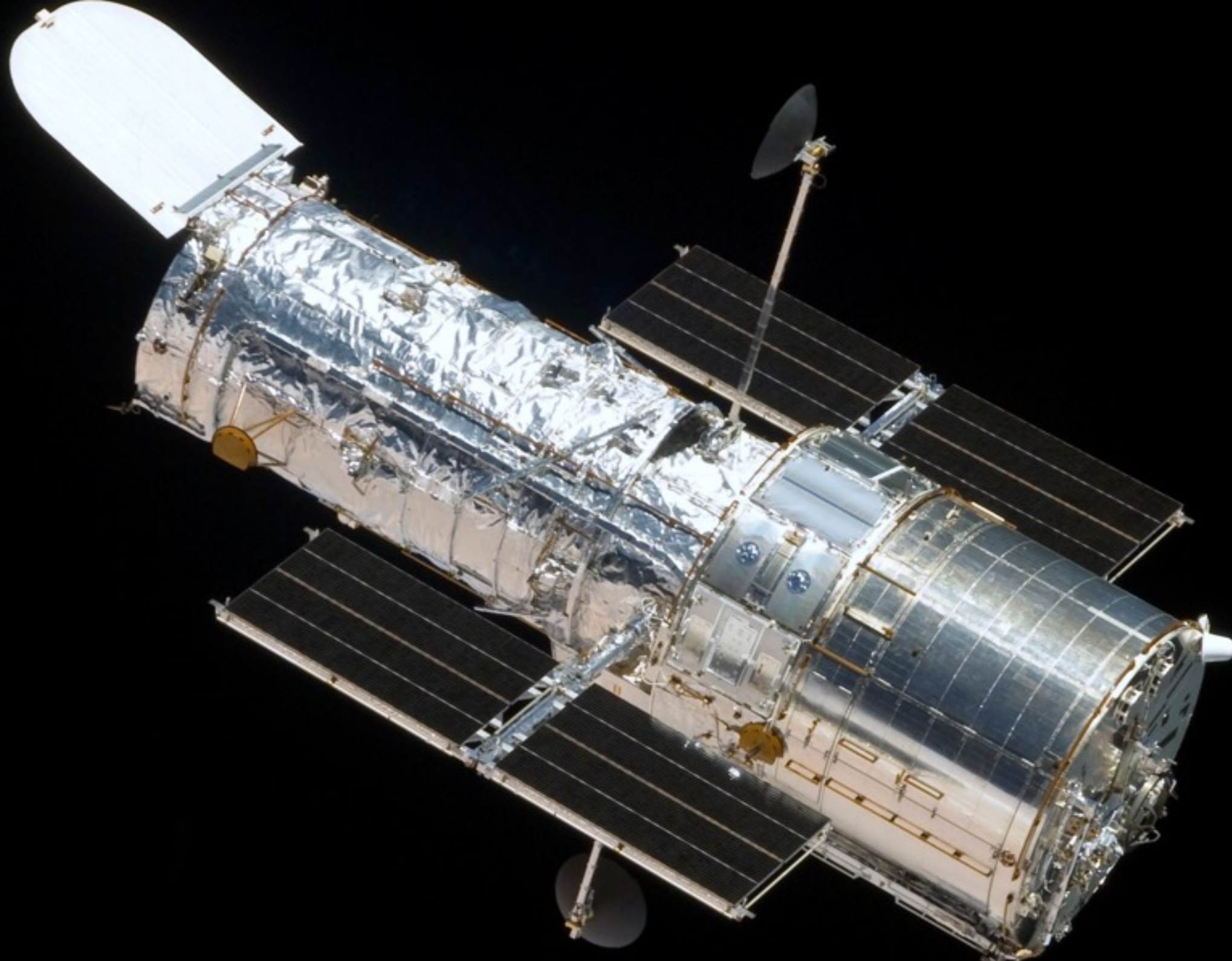
A report on RDA

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

Open Science and Open Data
Helsinki, Finland
1 October 2015



research infrastructure vs. e-infrastructure



“Research infrastructure
is stuff like telescopes”



"HST-SM4" by Ruffnax (Crew of
STS-125)



Overwhelming data
volume

Artist rendition courtesy SKA
Organisation



research infrastructure vs. e-infrastructure

a false dichotomy

e-Infrastructure *is* research infrastructure.

Modern research infrastructure *is* (or at least requires) e-Infrastructure.

It's about the *data*

Infrastructure is hard to conceive and describe because when it works, it's transparent, ubiquitous, and embedded in our daily work.











Dynamics of Infrastructure

Edwards, et al. 2007 Understanding Infrastructure: Dynamics, Tensions, and Design.



- Infrastructures become “ubiquitous, accessible, reliable, and transparent” as they mature.
- Systems  Networks  Inter-networks
 - “system-building, characterized by the deliberate and successful design of technology-based services.”
 - “technology transfer across domains and locations results in variations on the original design, as well as the emergence of competing systems.”
 - Finally, “a process of **consolidation characterized by gateways** that allow dissimilar systems to be linked into **networks**.”

Not what, but

When is infrastructure?

Not what, but

When and

Who is infrastructure?

Bridges and Gateways

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

— Edwards et al. 2007



Infrastructure is

Relationships, interactions, and connections
between people, technologies, and institutions

(that helps data flow and be useful)

Research Data Alliance



Vision

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

Mission

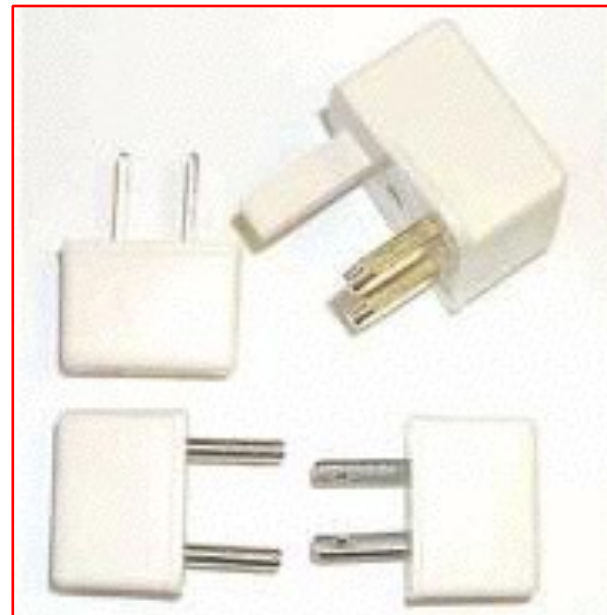
RDA builds the **social and technical bridges** that enable open sharing of data.

“Create - Adopt - Use”

(in 12-18 months)



Adopted Policy



Systems
Interoperability



Common Types,
Standards, Metadata



Sustainable Economics



Adopted Community
Practice



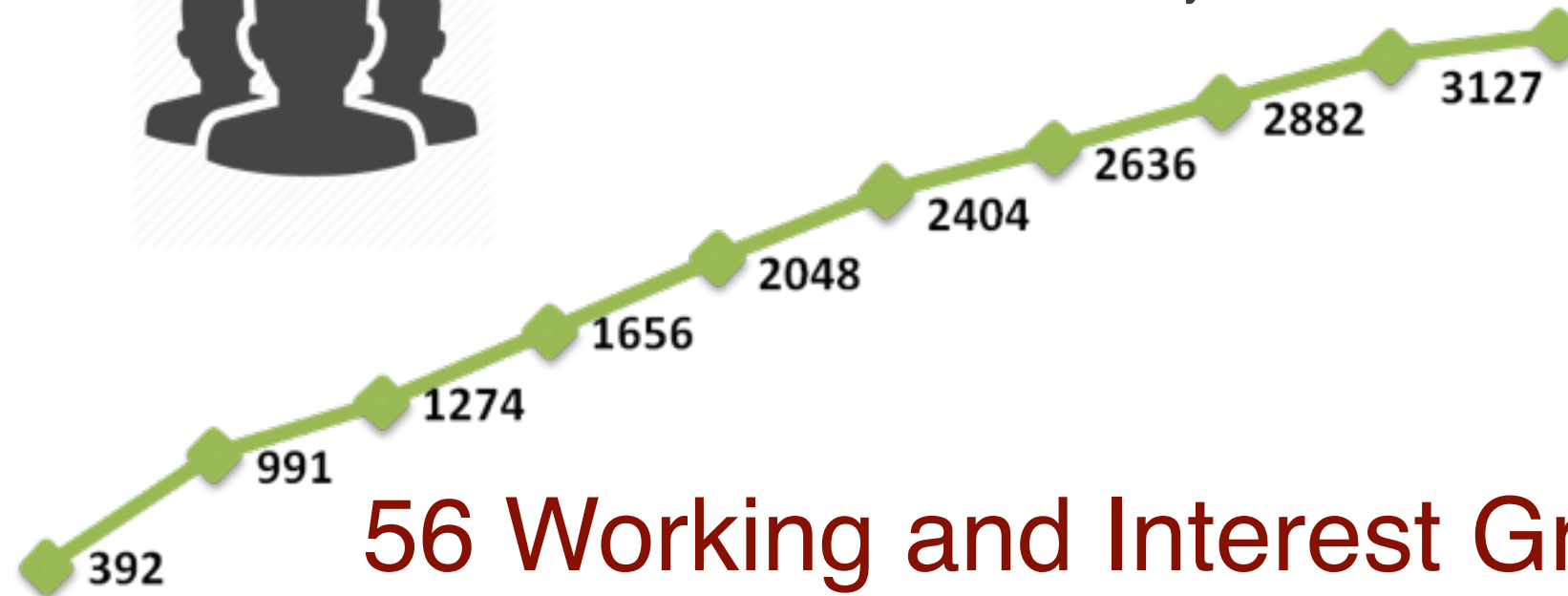
Training, Education,
Workforce

*Traffic Image:
Mike Gonzalez*

The Research Data Alliance Community Today



Total RDA Community Members: 3243

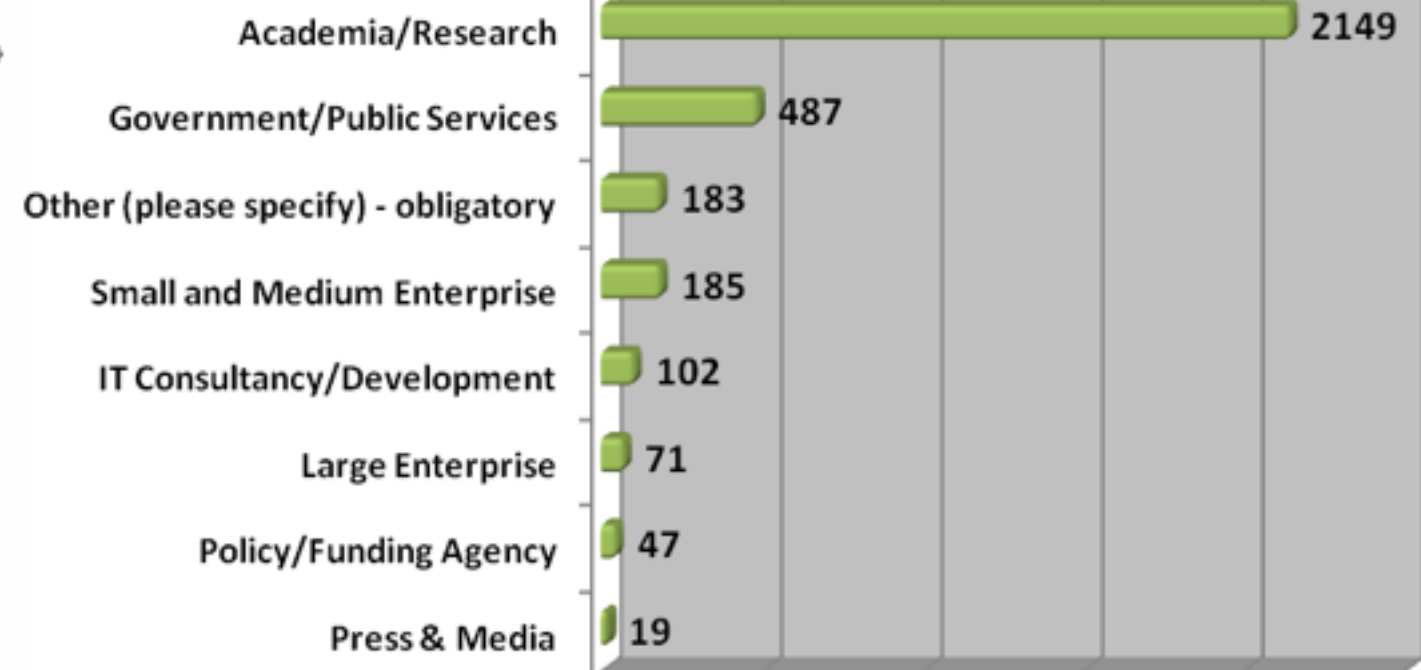
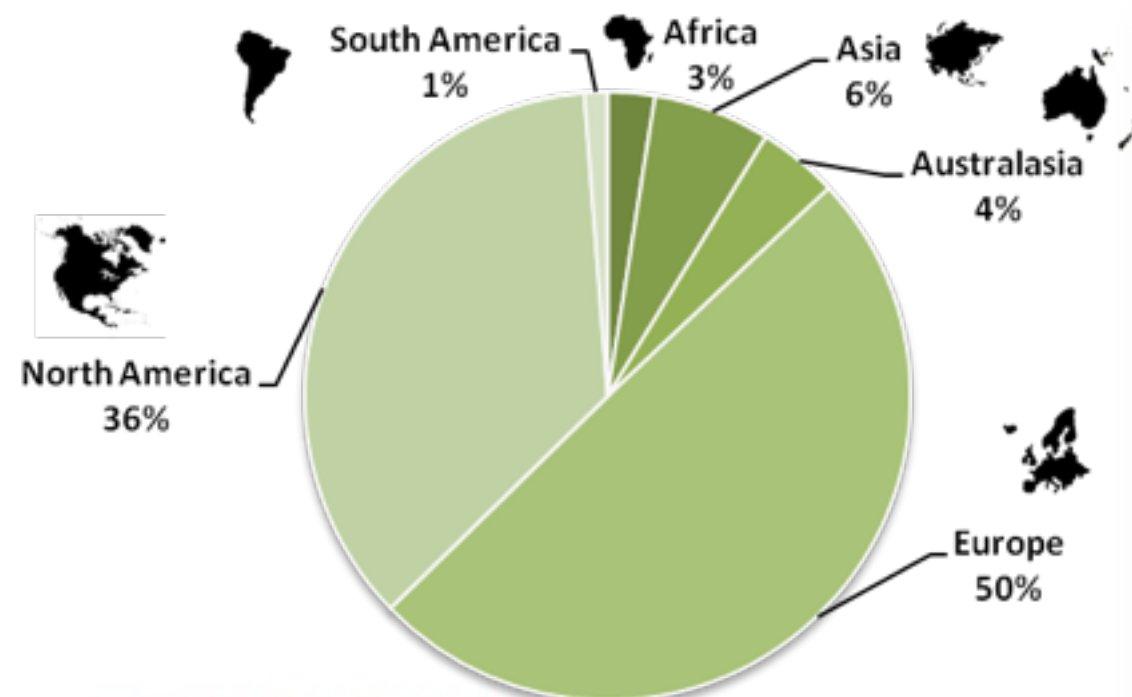


from 103 countries



56 Working and Interest Groups

May - July Aug - Oct Nov - Jan Feb - Apr May - July Aug - Oct Nov - Jan Feb - Apr May - July Aug - Sept



RDA Organisational Members and Affiliates



canarie

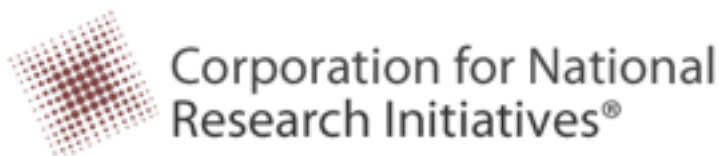


The Association
of Commonwealth
Universities



RatSWD

German Data Forum



NCSA



PURDUE
UNIVERSITY
LIBRARIES



RDC  DRC
Research Data Canada – Données de Recherche Canada



stm



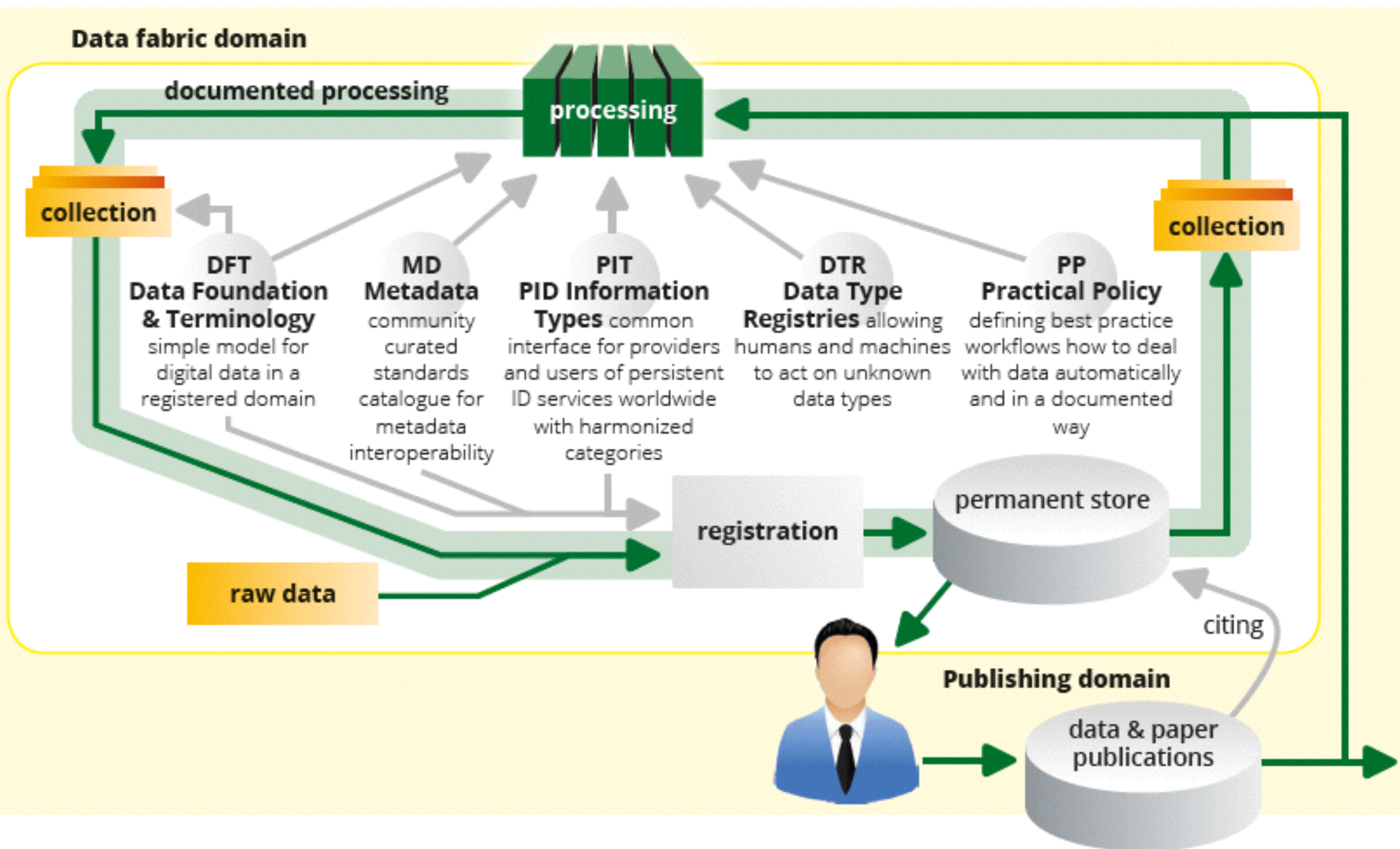
Initial Products—adopt one today!

- A basic vocabulary of **foundational terminology** and query tool to make sure we know what we're talking about.
- A **data type model and registry** ("MIME-types" for data) to help tools interpret, display, and process data.
- A **persistent identifier type registry** to help search engines understand what they are pointing to and retrieving.
- A basic set of **machine actionable rules** to enhance trust

Adopters presenting at P6

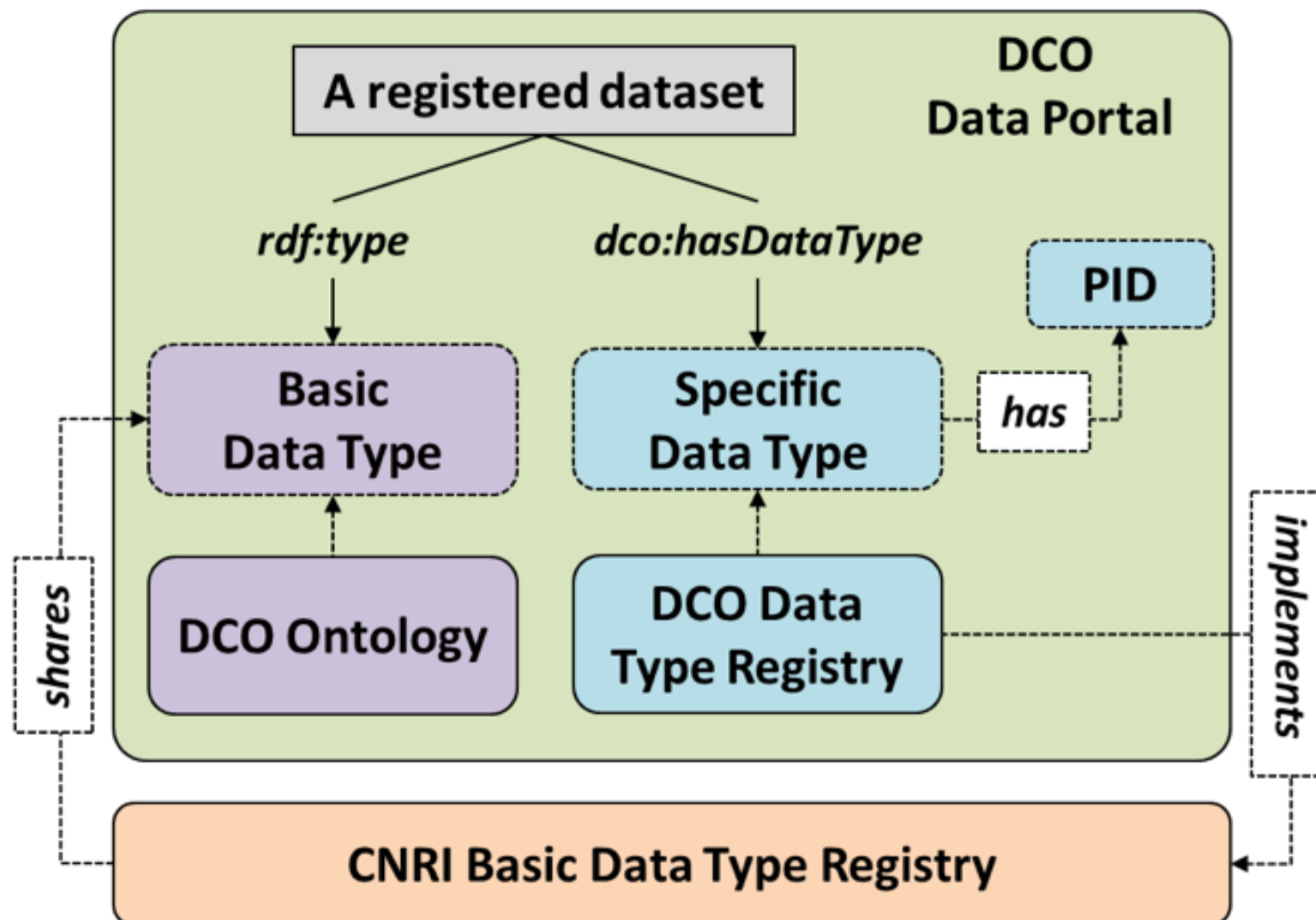
- *Deep Carbon Observatory,*
- *Platform for Experimental Collaborative Ethnography,*
- *Air Quality Community Catalog (datafed.net)*
- *Materials Innovation Infrastructure.*
- *EUDAT Collaborative Data Infrastructure,*
- *German Climate Computing Center (DKRZ)*
- *Common Language Resources and Technology Infrastructure (CLARIN).*

The “Data Fabric”





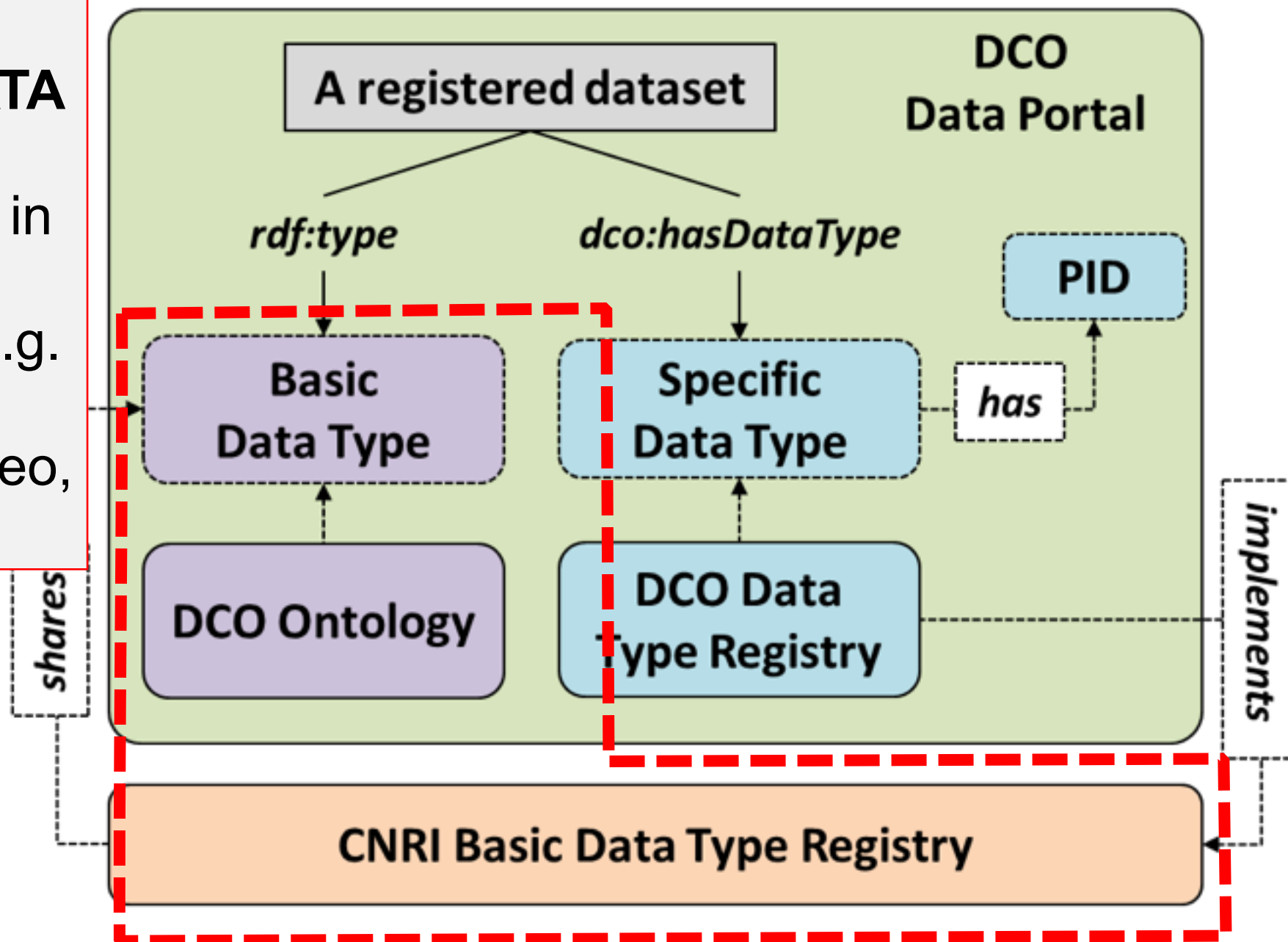
DTR in the Deep Carbon Observatory





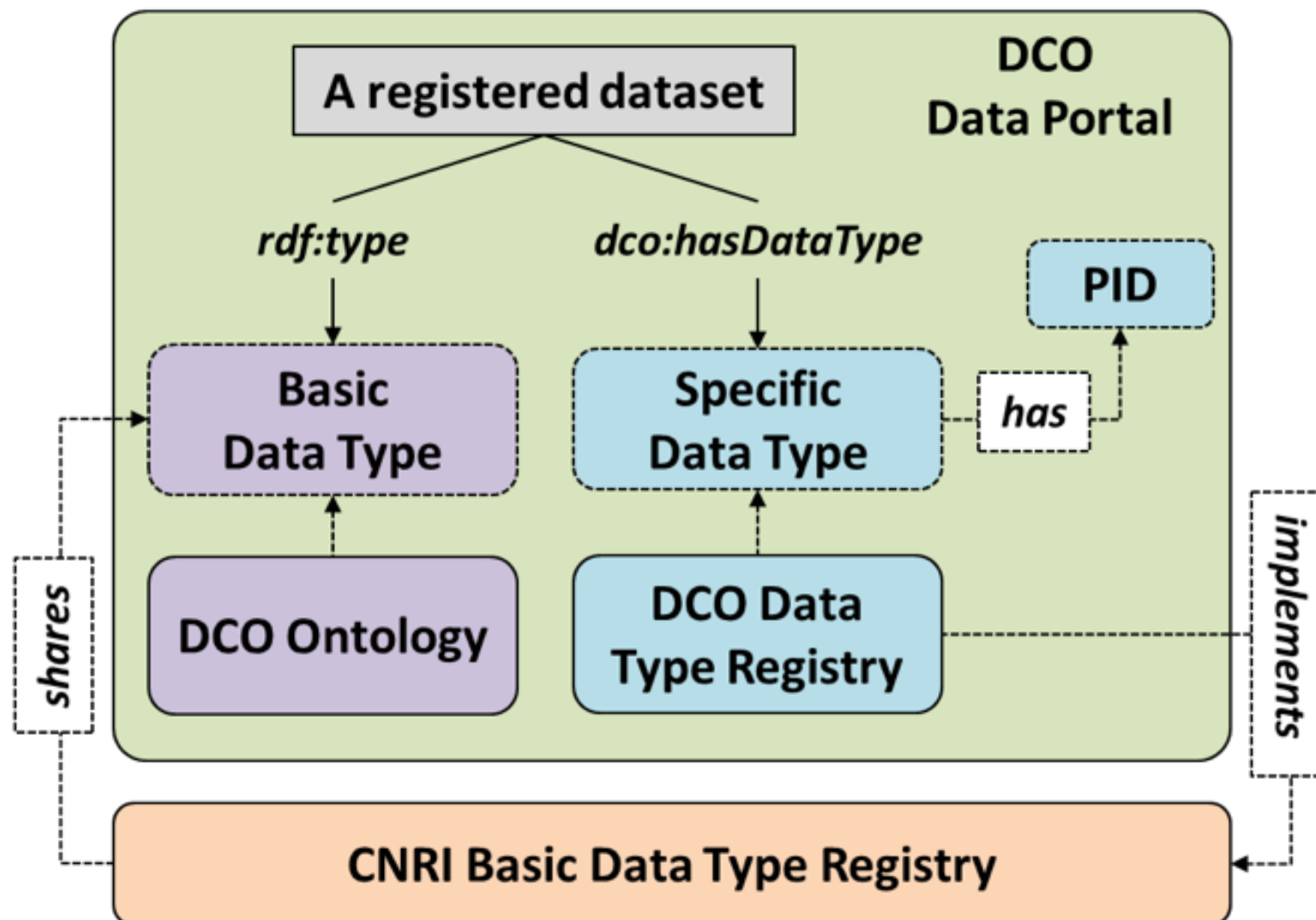
DTR in the Deep Carbon Observatory

The DTR primitives are comparable to a list of **BASIC DATA TYPE CLASSES** in the DCO ontology, e.g. Dataset, Image, Video, Audio, etc.





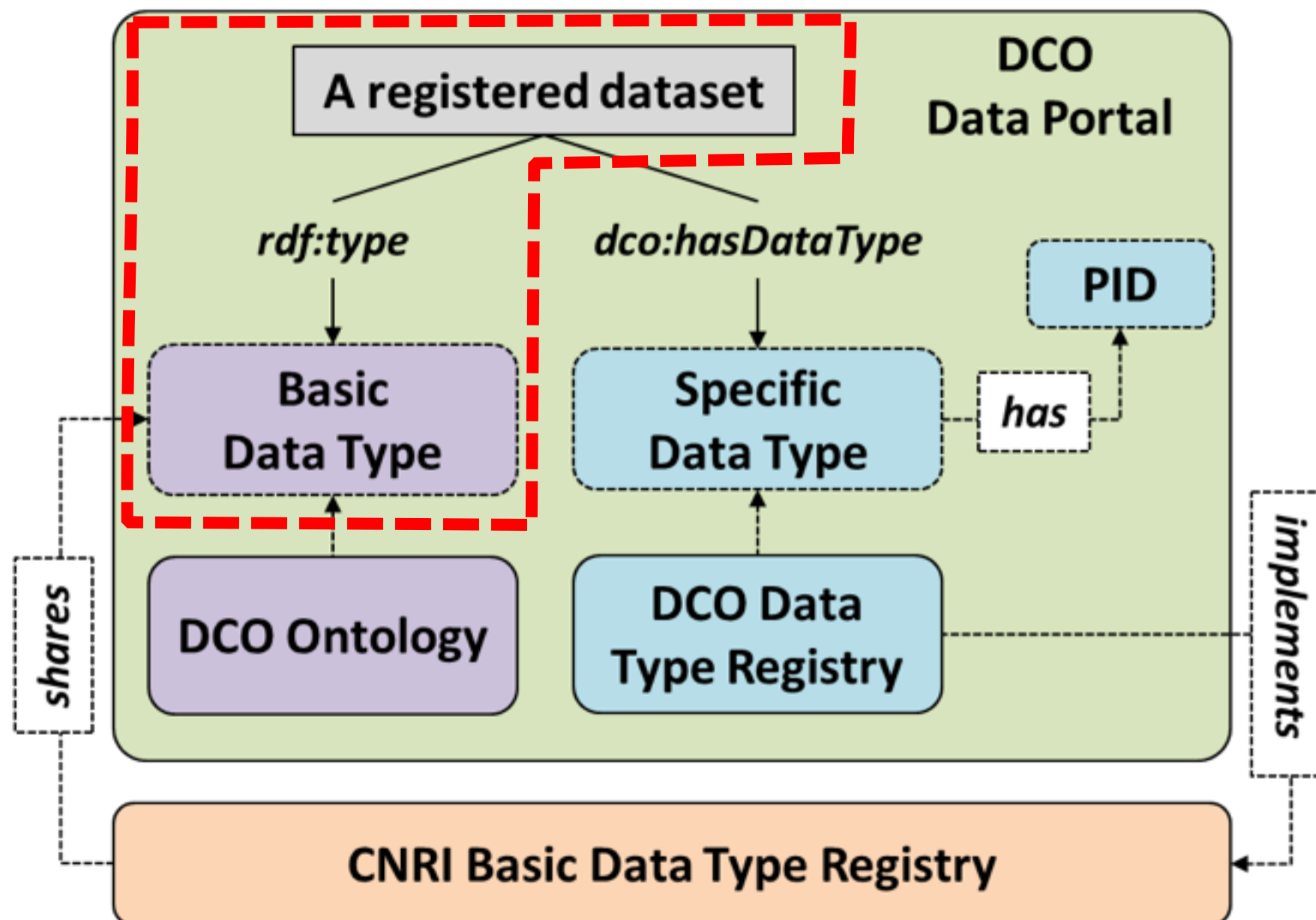
DTR in the Deep Carbon Observatory





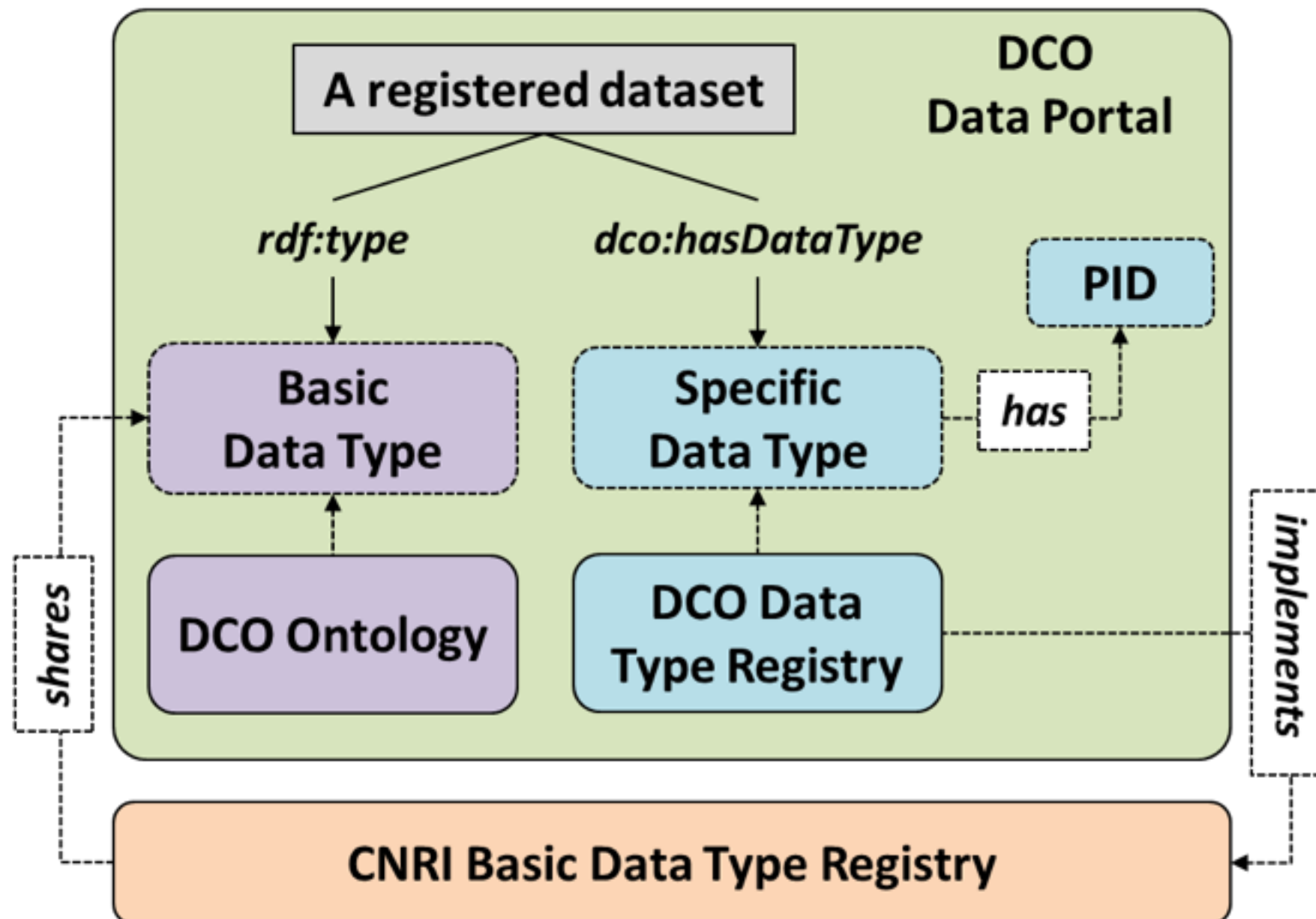
DTR in the Deep Carbon Observatory

A registered DCO dataset is asserted as an instance of one of those basic data type classes.



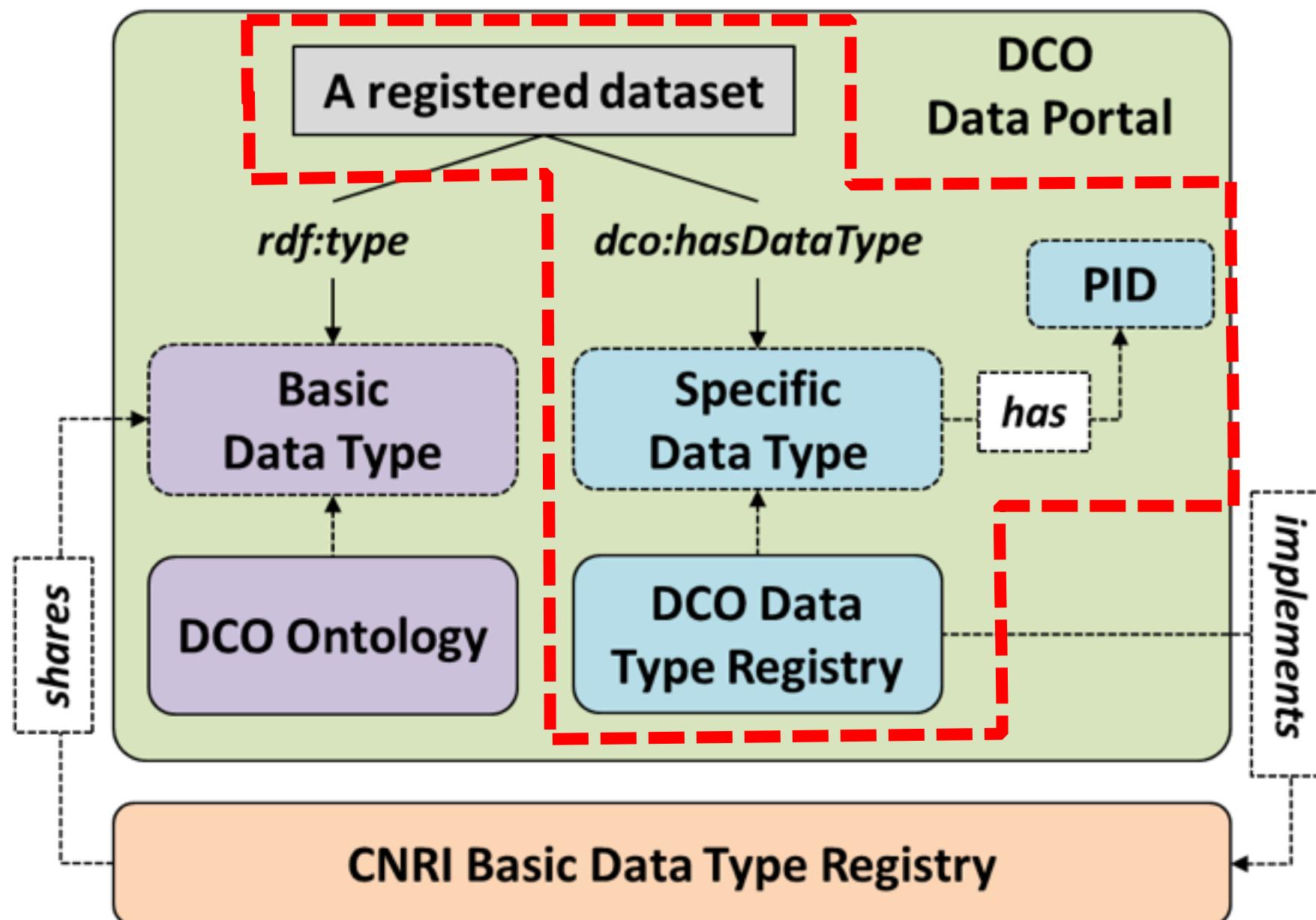


DTR in the Deep Carbon Observatory





DTR in the Deep Carbon Observatory



It is possible to further annotate the dataset with the **SPECIFIC DATA TYPES** defined within a DTR, and each data type has a unique PID.



Using Data Type as a facet in DCO dataset browser

► Years

► DCO Authors

► Communities

► Groups

► Projects

▼ Data Types

Comma-Separated Values (1) ⓘ

Keyhole Markup Language (1) ⓘ

Resource Description Framework (1) ⓘ

Thermodynamics of chemicals and minerals (1) ⓘ

1-2 of 2 < > 10 ▼

DCO Ontology
DCO ID: 11121/1490-1829-4848-8082-CC
Communities: Data Science Team
Groups: Data Science Community
Authors: Wang, Han; Erickson, John; Ma, Xiaogang (Marshall); Fox, Peter
Data Types: Resource Description Framework
Distributions:
DCO Ontology Distribution - 201404 (Direct access: DCO Ontology Turtle File - 201404)
DCO Ontology Distribution - 201408 (Direct access: DCO Ontology Turtle File - 201408)
DCO Ontology Distribution - 201405 (Direct access: DCO Ontology Turtle File - 201405)

The Heat Capacities of Magnesium, Zinc, Lead, Manganese and Iron Carbonates at Low Temperatures
DCO ID: 11121/4589-9713-2502-3685-CC
Communities: Data Science Team; Extreme Physics and Chemistry Community
Groups: Data Science Community
Authors: Zhong, Hao; Ma, Xiaogang (Marshall); Ghiorso, Mark
Data Types: Thermodynamics of chemicals and minerals
Distributions:
2015-03-22-1532

Thermodynamics of chemicals and minerals | [Data Type](#)DCO ID [11121/9177-8600-7213-5328-CC](#)[Overview](#) | [Identity](#) | [Other](#) | [View All](#)**expected uses**

The data type is for thermodynamics of chemicals and minerals. Records cover two major topics: **Enthalpy & Entropy**. Detailed items include but are not limited to:

- Mineral Name
- Molecular Formula
- Molecular Weight
- Temperature (T, °K)
- Temperature Change
- Heat Content (Cp, calorie/mole)
- Entropy Increment
- Data Source

Additionally, information about materials used in the test of thermodynamic features can also be recorded. Detailed items can include:

- Material
- Source
- Sample, g.
- Density Density temp., °C
- Purity, %
- Impurities
- Impurities corrected for

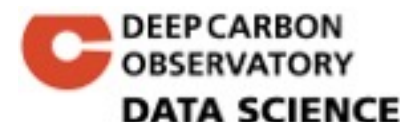
has parameter[Data Source](#)



Conclusions

- The methodology of RDA DTR and PIT is highly implementable, especially in the environment of the Semantic Web.
- The technical framework in the current demonstration systems of DTR and PIT can be adapted or further extended for production uses.
- Initial good researcher response (they recognize their data types)
- Slides with backup detail at: http://tw.rpi.edu/web/doc/20150921_slides_RDA_P6.pptx
- Contact Marshall at max7@rpi.edu

Thank you!



Project Overall:

- Significant technical modifications to the AQComCat were completed within the 8 month grant period
 - Cosmetic improvements would still need to be undertaken to optimize the revised AQComCat
- To make the implicit knowledge of the domain researchers explicit for the non domain researcher requires significant generalization and explanation of the terms and could be improved upon

Adoption Overall:

- Adopting as an outsider is possible!
- Working groups were available to discuss outcomes and implementation at length

Research Data Alliance Practical Policies for Data Management Adoption

Platform for Experimental, Collaborative Ethnography (PECE) RPI, Troy, NY, USA

Luis Felipe R. Murillo
(Berkman Center for Internet and Society)

Research Data Alliance – 6th Plenary
Paris, September, 23, 2015

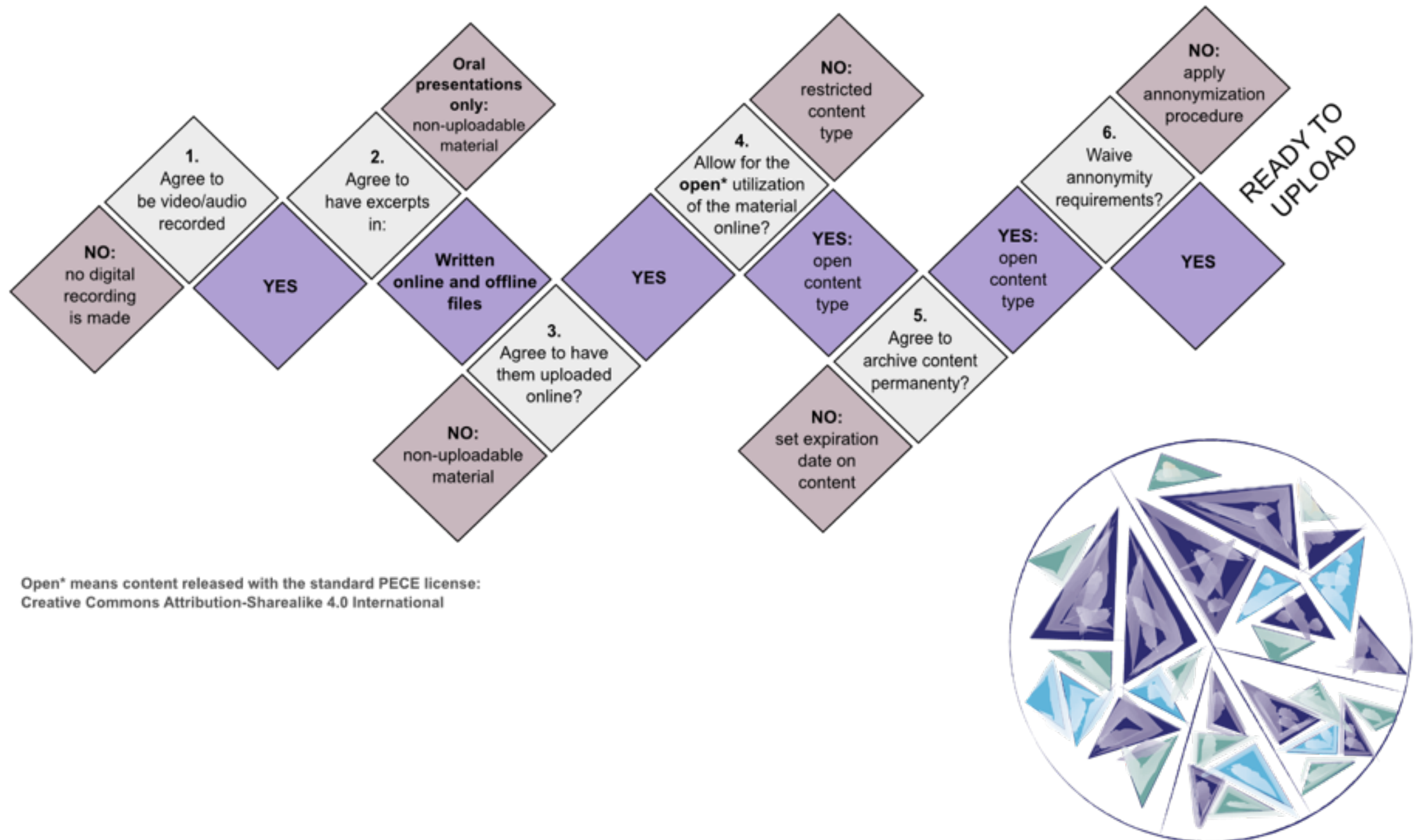


PECE Data Management

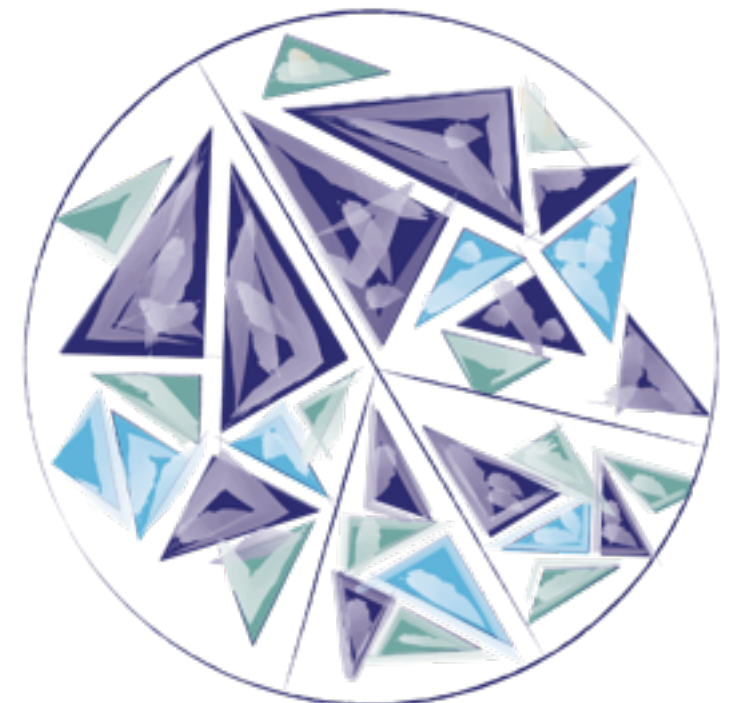
- **Contextual Metadata Extraction:**
PECE Open API
- **Data Access Control:** *role-based permission system and permissions per digital object*
- **Data Format Control:** *only open formats and standards are allowed in the platform*
- **Backup:** *automated, redundant, and encrypted*
- **Restricted Searching:** *based on user roles, easy to remove a digital object from the search index, distributed ElasticSearch*
- **User Agreements:** *prepared with the Cyberlaw clinic of the Berkman Center at Harvard*
- **And more...**



Institutional Review Board (IRB) informed consent form questions translated into PECE permission system



Open* means content released with the standard PECE license:
Creative Commons Attribution-Sharealike 4.0 International



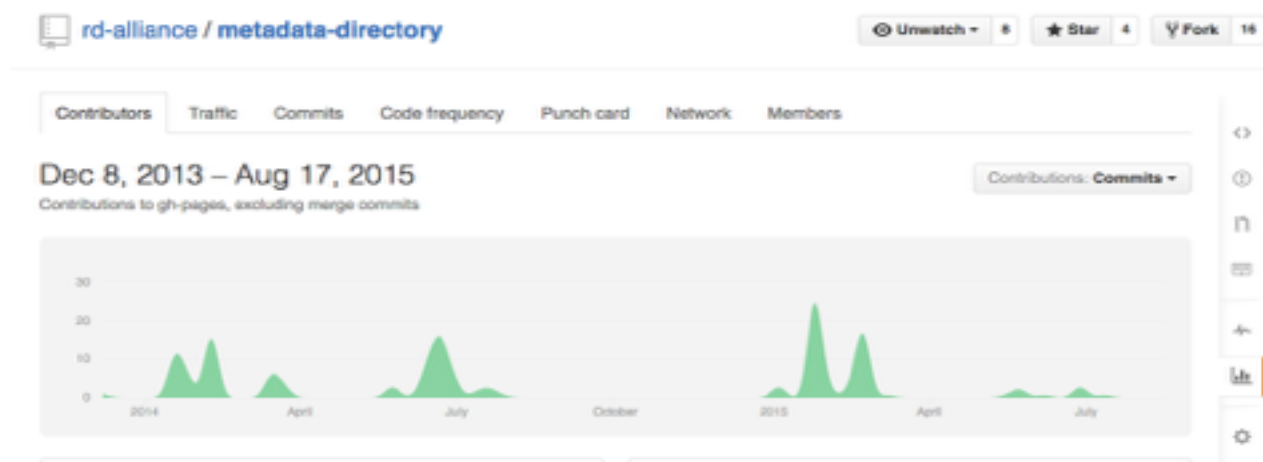
New Products — adopt one today!

- A **metadata standards directory** so we can describe similar things consistently
- A **dynamic-data citation methodology** so we can reference precise subsets of changing data.
- Semantically linked **terms describing wheat data** so we can share harvest and related information around the world
- **Services and methods for finding data across multiple registries**, to help cross disciplinary and multi-faceted discovery.

- Metadata Standards Directory WG
 - 135 members
 - Representing multiple disciplines (Environmental Science, Geology, Bioinformatics, Libraries, Computer Science, etc.)
 - Multiple countries represented (US, UK, Finland, Italy, France, etc.)
- Deliverables:
 - Directory of descriptive, discipline-specific metadata standards to:
 - Promote the discovery, access and use of standards
 - Improve the state of research data interoperability and reduce duplicative standards development work
 - Expanded and updated the DCC Metadata Catalogue
 - Website to add or correct standards
 - Collection of use cases

Endorsements/Adopters

- Who are the adopters and how have they used the deliverable?
 - UK Digital Curation Centre (DCC)
 - 18,339 page views 1 January -16 September 2015
 - Data Observation Network for Earth (DataONE)
 - Included in Best Practices Database
 - ~16,500 users/quarter and ~20,250 sessions/quarter
- GitHub Use to Update DCC Directory



- Reiterate who could use this deliverable
 - Researchers to find appropriate metadata to make their datasets available, discoverable, interoperable, and curatable
 - Data managers / librarians for creating local standards for researchers in their jurisdiction
 - Go beyond just 'dumping' the metadata specification
 - Requires contextual metadata to explain context in which it can be used
 - And appropriate scripts for downloading/implementing and APIs for interoperation

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned (aka history)

Researcher creates working-set via some interface:

- Access → **assign PID to QUERY**, enhanced with
 - **Time-stamping** for re-execution against versioned DB
 - **Re-writing** for normalization, unique-sort, mapping to history
 - **Hashing** result-set: verifying identity/correctness

leading to landing page

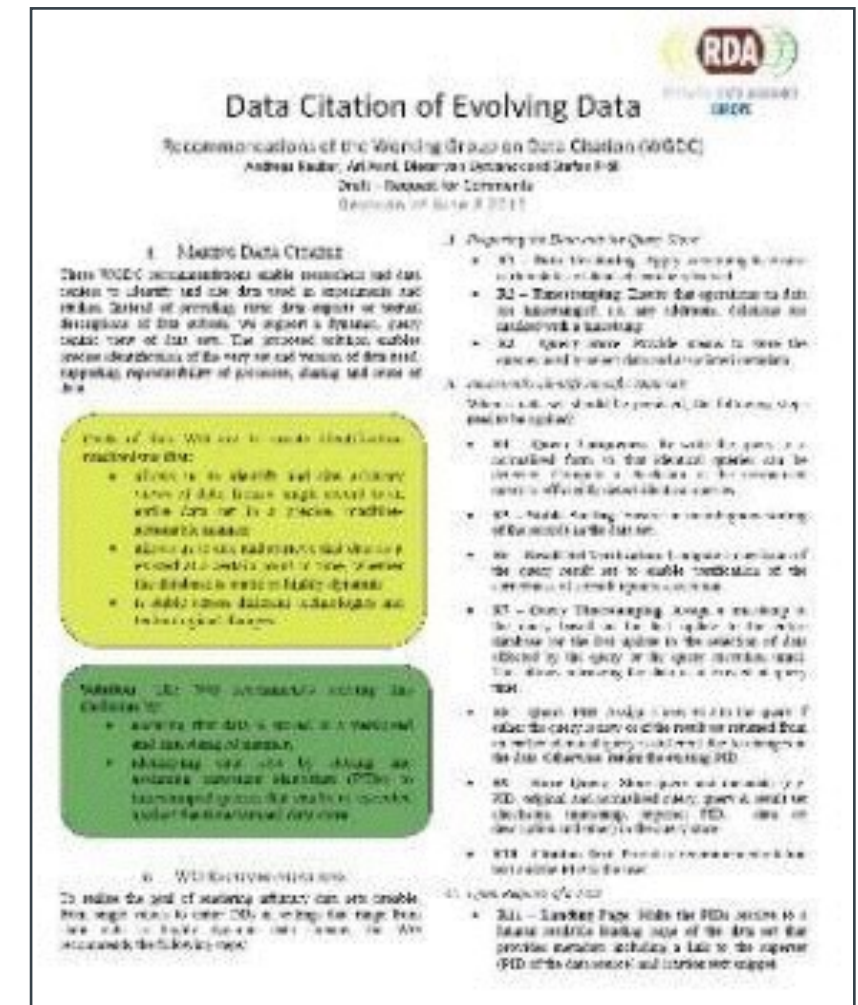
S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation**. In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013

http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Output / Results

<#>

- 14 Recommendations grouped into 4 phases:
 - Preparing data and query store
 - Persistently identifying specific data sets
 - Resolving PIDs
 - Upon modifications to the data infrastructure
- 2-page flyer
- Technical Report to follow
- Reference implementations (SQL, CSV, XML)
- Pilots



- Pilots and implementations by
 - LNEC: Critical Infrastructure Monitoring System
 - Virtual Atomic and Molecular Data Centre
 - NERC (UK Natural Environment Research Council Data Centres)
 - ARGO Buoy Network
 - River Flow Dataset
 - ESIP (Earth Science Information Partners)
 - BCO-DMO
 - DEXHELPP – Social Security Data
 - ENVRIplus: Carbon Observation System
 - Million Song Database, IR Benchmark DBs
 - Several others under discussion...



The Wheat Data Interoperability WG

- Aims: contribute to the improvement of Wheat related data interoperability by
 - Building a common interoperability framework (metadata, data formats and vocabularies)
 - Providing guidelines for describing, representing and linking Wheat related data

Contributors



Sponsors



Active members: Alaux Michael (INRA, France), Aubin Sophie (INRA, France), Arnaud Elizabeth (Bioversity, France), Baumann Ute (Adelaide Uni, Australia), Buche Patrice (INRA, France), Cooper Laurel (Planteome, USA), Fulss Richard (CIMMYT, Mexico), Hologne Odile (INRA, France), Laporte Marie-Angélique (Bioversity, France), Larmand Pierre (IRD, France), Letellier Thomas (INRA, France), Lucas Hélène (INRA, France), Pommier Cyril (INRA, France), Protonotarios Vassilis (Agro-Know, Greece), Quesneville Hadi (INRA, France), Shrestha Rosemary (INRA, France), Subirats Imma (FAO of the United Nations, Italy), Aravind Venkatesan (IBC, France), Whan Alex (CSIRO, Australia)

Co-chairs: Esther Dzalé Yeumo Kaboré (INRA, France), Richard Allan Fulss (CIMMYT, Mexico)

The deliverables

- Guidelines (<http://wheatis.org/DataStandards.php>)
 - Data exchange formats
 - Example: VCF (Variant Call Format) for sequence variation data, GFF3 for genome annotation data, etc.
 - Data description best practices
 - Consistent use of ontologies, consistent use of external database cross references
 - Data sharing best practices
 - Share data matrices along with relevant metadata (example: trait along with method, units and scales or environmental ones)
 - Useful tools and use cases that highlight data formats and vocabularies issues
- A portal of wheat related ontologies and vocabularies (<http://agroportal.lirmm.fr/ontologies?filter=WHEAT>)
 - Allows the access to the ontologies and vocabularies through APIs.
- A prototype
 - Implementation of use cases of wheat data integration within the AgroLD (Agronomic Linked Data) tool: <http://volvestre.cirad.fr:8080/agrold/>

Benefits for many target users

- For data managers, data providers
 - One stop shop for relevant information related to data management → arise awareness, avoid duplicated efforts, foster adoption of common practices
 - Facilitate the use of common data exchange formats → easy data sharing/submission to international repositories
 - Foster a standardized description of datasets with consistent use of ontologies and metadata → increase the identification, the findability and the usability of the datasets
- For data scientists, bioinformaticians
 - Facilitate the access, integration and analysis of data from various sources
 - Access to data of higher quality
- For top management, researchers
 - Increase the chance to answer complex questions

- **DDRI Participants:** ANDS, Dryad, CERN, DataPASS, da-ra, Thomson Reuters, VIVO Cornel, DANS, DataCite and Data Curation Unit (DCU)
- **Deliverable:** a proposed model for connecting datasets on the basis of co-authorship or other collaboration models such as joint funding and grants.

The proposed model has been adopted by ANDS and implemented as **Research Data Switchboard** – an open source software platform.

Use Cases:

- Repositories: finding connected datasets across multiple platforms
- Universities: finding datasets by their researchers
- Researcher: finding similar datasets connected by co-authorship and joint funded grants

- Australian National Data Service
 - <http://rd-switchboard.net>
- NCI - National Computational Infrastructure
 - Connecting Australian research data across multiple platforms
- University of Sydney (Australia)
 - Connecting datasets by the researchers from the University of Sydney

Next Products—coming next Plenary!



- A **unified repository certification scheme** to reduce confusion and improve trust.
- A suite of **data publishing-related services** for
 - measuring bibliometrics
 - managing data workflows
 - interconnecting articles and data

WDS/DSA Common Certification Requirements and Procedures

- WG began with 4 members from WDS and 4 from DSA with support from the WDS-IPO but has expanded to include several external members; WDS with Earth and Space Sciences background and DSA with Humanities and Social Sciences; from Europe, USA, China, South Africa, Australia, (Japan).
- The deliverable
 - Common Basic certification requirements/criteria
 - Implementation plan for Common Procedures
 - Testbed – “Real-world” valuation of Common Requirements and Procedures

Impact of the Deliverable

- Who is impacted as a result of this deliverable
 - Include potential scenarios demonstrating acceleration of innovation (cross-domain collaboration), time savings, economic savings, etc.
- Will provide a step towards having more coherent, increasingly stringent and compatible standards for repository certification
- DSA–WDS certification standard adoption will create a critical mass of certified repositories across a range of domains and disciplines
- Data Collectors, Funders, Publishers and Users – deliverable inspires trust, which is at the heart of sharing and archiving data

- Who are the adopters and how have they used the deliverable?

ICSU World Data System

Data Seal of Approval

- Common **L**anguage **R**esources and Technology **I**nfrastructure (CLARIN)
- IOC International Oceanographic Data and Information Exchange (IODE) programme

Other repositories

.....

- Working Group structure
 - 52 members, mostly European, USA
 - Mixture of academics, repository managers, academic publishers
- “Next steps for Bibliometrics for Data” will be based on:
 - WG survey results (presented P4 and P5)
 - Spreadsheet of metrics being collected by repositories - Still open for contributions! <http://bit.ly/1MpyW4K>
 - Shared results from other projects – understanding the challenges and answering the questions posed in the case statement
 - Preliminary analysis of data DOI resolutions
 - Supporting and evaluating tools from other projects
 - Preliminary guidance for the community - “minimal” rather than “best” practice – get people discussing the issues and coming up with solutions!

- Who is impacted as a result of this deliverable?
 - Repository managers, librarians, academic publishers ...

Other projects in this space:

- CASRAI data level metrics
- PLOS Making Data Count
- NISO altmetrics
- Jisc Giving Researchers Credit for their Data
- Lots of interest in this area!
- Mapping the bibliometrics for data landscape

- Who are the adopters and how have they used the deliverable?
 - None yet, as deliverable isn't finalised!
 - Anticipating repository managers, academic publishers, researchers, funders, research institutes, librarians.

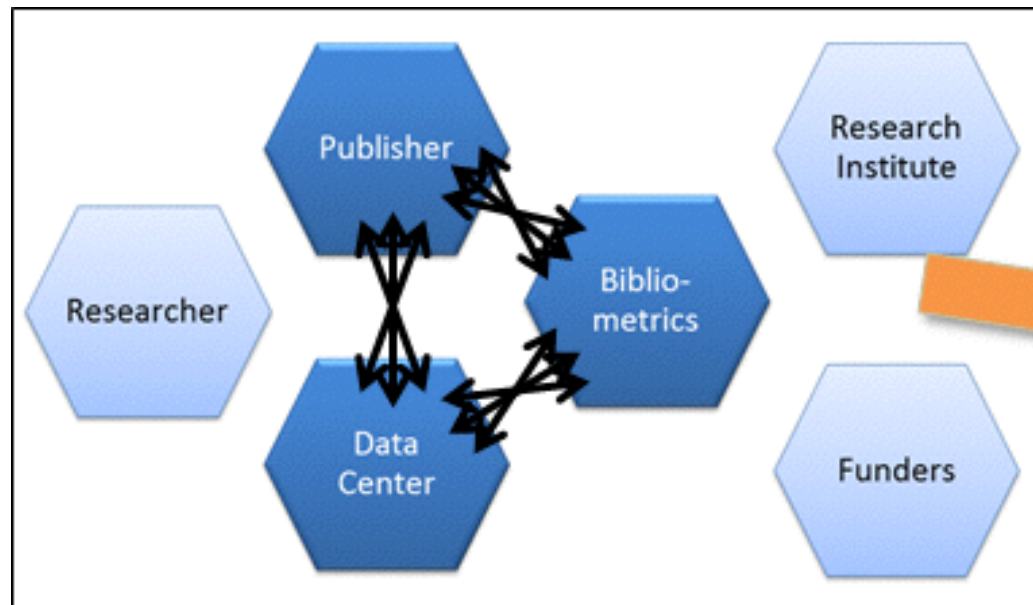
Data Publishing Models

- Workflows -

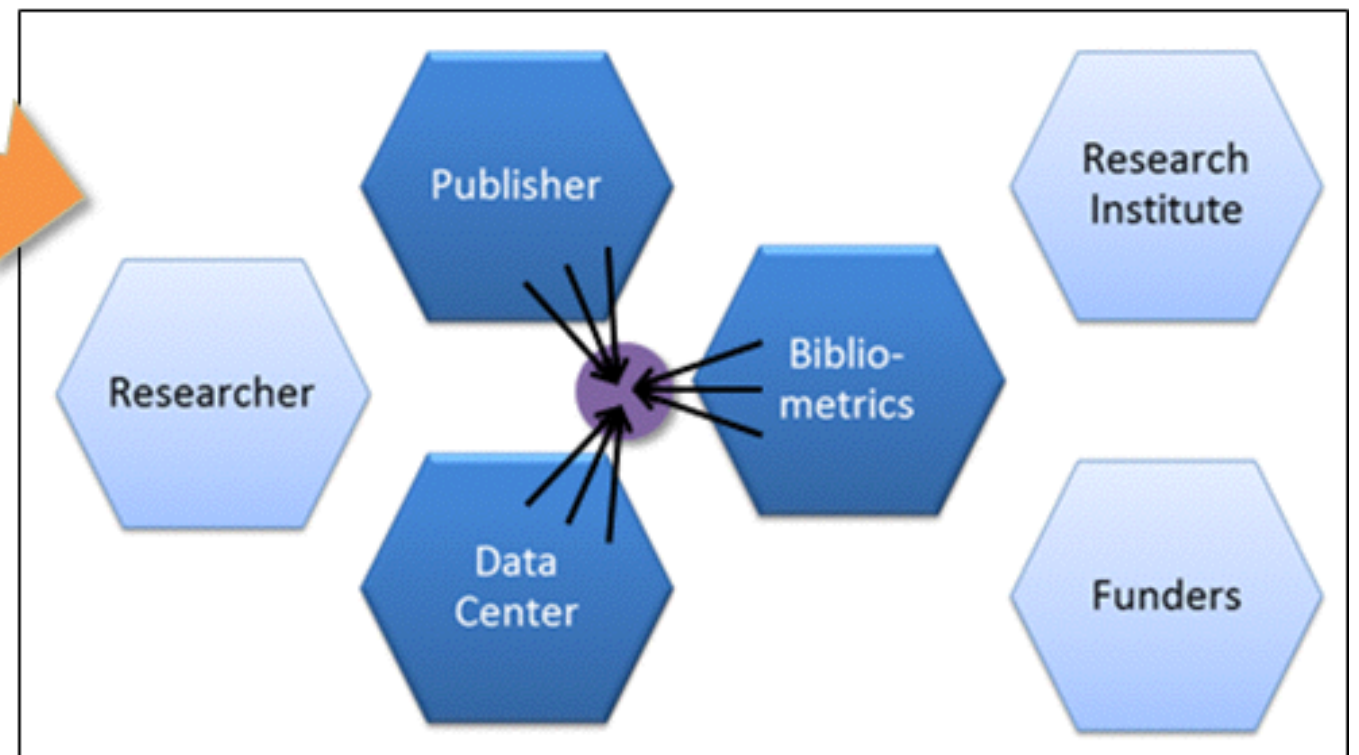
- First comprehensive review of current data publishing
 - Data repositories and data journals
 - Building blocks, best practices
- Recommendations for trusted data publishing
- Reference models for researchers, repositories, publishers who wish to publish data
 - Including links/recommendations to/for shared/open sources tools to re/use for individual workflow components
 - “beginner’s guide”

Introducing the Data Publishing Services WG

How to move from a plethora of (mostly) bilateral arrangements to a one-for-all service model infrastructure for the research data publication landscape?



- Increase interoperability
- Decrease systemic inefficiencies
- Power new tools and functionalities to the benefit of researchers



Main deliverable: an open, universal cross-linking service ⟨#⟩

Primary Focus: Universal literature – data cross-linking service

- Given article A, what relevant data D exists – and vice versa
- Additional metadata about the nature of the relationship, e.g. supplementary data, related data, formal citation.
- Additional metadata for article and/or data set

Why? Improve visibility, discoverability, re-use and reproducibility

Many organizations are already doing this – for example:

- Data repositories keep track of articles that cite, or refer to, their data
- Publishers devise applications to link the articles they publish with data hosted externally
- Providers of bibliographic information (i.e. repositories) are increasingly looking at data alongside the traditional article output
- Organizations such as CrossRef, DataCite and OpenAIRE are developing systems to track or infer relationships between data and the literature

But it's all very scattered, which limits the value!

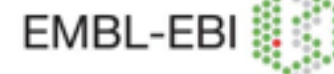
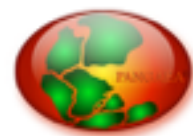
Linking data and the literature

<#>

Deliverable: We want to bring existing article/data links together, normalize them using a common schema, and expose the full set as an open service

Philosophy:

- Open, collaborative & inclusive, cross-stakeholder
- Domain-agnostic, (aim for) comprehensiveness
- Quality and provenance is key
- Flexible, test & learn, hands-on



Plenary 7

1-3 March 2016

Tokyo, Japan





Info:

enquiries@rd-alliance.org

[@resdatall](https://twitter.com/resdatall)

research data sharing without barriers
rd-alliance.org