RDA ICT Technical Specifications

-

Creating positive impact on research and business

The first four - DFT, PIT, DTR & PP

Raphael Ritz, MPCDF & RDA Europe

RDA EU Data Innovation Forum

Brussels, January 30, 2018

ICT technical specifications - Definitions

- Technical Specification
  - A document that prescribes **technical requirements** to be fulfilled by a product, process, service or system
- Standard
  - A technical specification, **adopted by a recognised standardisation body**, for repeated or continuous application, with which compliance is not compulsory
- ICT Technical Specification
  - A technical specification in the field of information and communication technologies
- **Identified ICT Technical Specifications**
  - Can be referenced in public procurement, primarily to enable interoperability between devices, applications, data repositories, services and networks.
  - Official status under the EU public procurement legislation: “Common Technical Specification”
  - Comply with Regulation No 1025/2012, Annex II

Identification of ICT technical specifications : a lighter procedure than standardisation

Identification of ICT specifications in Europe

The European Commission has a flexible approach to standardisation when identifying new ICT technical specifications.

WHY?

The European Commission can identify ICT technical specifications that are not national, European, or international standards, provided they meet precise requirements. Once identified and approved, these specifications can then be referenced in European public procurement. This flexible approach allows the EU to ***respond to the fast evolution of technology in ICT****. It also helps encourage competition, promote interoperability and innovation, and facilitate the provision of cross-border services.*

The Research Data Alliance has presented 9 of the RDA Recommendations to be evaluated and acknowledged as ICT Technical Specifications.

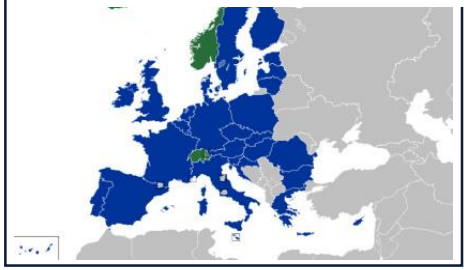
Who is involved in this process?

- The European **Multi Stakeholder Platform** (MSP) is an expert **advisory group** on ICT standardisation.
- It deals with:
  - Potential future ICT standardisation needs in support of European legislation, policies and public procurement;
  - Technical specifications for public procurements, developed by global ICT standards-developing organisations;
  - Cooperation between ICT standards-setting organisations;
  - The Rolling Plan, which provides a multi-annual overview of the needs for preliminary or complementary ICT standardisation activities in support of the EU policy activities

The Multistakeholder platform (MSP) is chaired and coordinated by the European Commission.

MSP Members

Member States and EFTA countries



ICT Standardisation Bodies

CEN
CENELEC EUROPEAN COMMITTEE FOR ELECTROTECHNICAL STANDARDIZATION
ETSI Celebrating 25 years
IEC
ISO
ITU
ecma INTERNATIONAL
OMG WE SET THE STANDARD
IEEE Advancing Technology for Humanity
UN/CEFACT SIMPLE, TRANSPARENT AND EFFECTIVE PROCESSES FOR GLOBAL BUSINESS
GS1

Industry, SMEs and society representatives

Inria
EDiMA Shaping Internet Policy in the EU
ecta
BUSINESSEUROPE
Openforum europe open, competitive choice for IT users
Small Business Standards
ORGALIME
AGE Platform Europe
ANEC
etno
Internet Society
ECOS
GSMA
EBU
EuroISPA
european disability forum
ECIS European Committee for Interoperable Systems
OPEN INTERNET PROJECT
Cable Europe
industriAll EUROPEAN TRADE UNION
DIGITALEUROPE

The MSP is composed of ICT standard experts

RDA Compliance with Requirements for ICT Technical Specifications

ANNEX II: Requirements 3 (a) (b) (c)

Openness: RDA WG processes & procedures are public & completely open

Consensus: foundation upon which RDA is built

- All processes and procedures, in connection with focus, work plans, deliverables, milestones & tangible specifications / recommendations are **consensus** based.
- WG work plans, activities & outputs go through an **open** & **transparent** public community review process in addition to feedback provided by RDA Technical Advisory Board and Council.

Transparency: all information available, balance and harmonisation ensured and all feedback considered and responded to.

RDA Compliance with Requirements for ICT Technical Specifications

ANNEX II: Requirements 4 (a) (b) (c) (d) (e) (f)

- ✓(a) maintenance
- ✓(b) availability
- ✓(c) intellectual property rights
- ✓(d) relevance
- ✓(e) neutrality and stability
- ✓(f) quality

MAIN TARGET MARKET:

- Service Providers
- Data Providers
- Repositories
- E- & Research Infrastructures
- SW Developers
- Data Scientists
- Researchers & Scientists

Open information exchange & **involvement of all interested categories** are at the core of RDA's vision of an open, global, collaborative science. RDA adopts a **consultative approach** involving all **relevant actors** to spur international collaborations necessary to address the global challenges.

Why RDA ICT technical specifications?

RDA Technical specifications

- **Data federation cost efficiency**
- **Avoid / reduce technology & market lock-in**
- **Innovation friendly**
- **Open & User-driven**
- **Enable European and Global data Interoperability**
- **Increased implementation due to Public procurement**

e-Infrastructure approach

Service orientation

federation, virtualisation

Multiple funding sources

flexible, agile business models

Innovation

User and technology-driven

Interoperability of data and computing

RDA Working Groups – Technical Specification Producers

RDA Working groups (WGs) accelerants to advance global data-driven discovery, interoperability & innovation in the long-term.

1. Case Statement open for public comment
2. Comments integrated -> *Revised case statement*
3. Work begins (12-18 month duration) & presented every 6 months
4. Outputs released for public comment (RfC)
5. Outputs revised -> integration / modification
6. Endorsed & Recognised by RDA Council
7. Openly available for implementation / adoption

Testing & Implementation

RDA & first 4 Technical Specifications approved

PUBLISHED IN OFFICIAL EC JOURNAL JULY 2017

✓ TS1 **The Basic Vocabulary of Foundational Terminology and Query Tool** - produced by the Data Foundation & Terminology WG which ensures researchers use a common terminology when referring to data.

✓TS2 **The Data Type Model and Registry** published by the Data Type Registries WG providing machine-readable and researcher-accessible registries of data types that support the accurate use of data

✓TS3 **The Machine Actionable Policy Templates** produced by the Practical Policy WG designed to support data sharing and interchange between communities.

✓TS4 **The Persistent Identifier Type Registry** produced by the PID Information Types WG, a conceptual model for structuring typed information to better identify PIDs, common interface for access to this information.

Data Foundation & Terminology: Motivation

Bob Kahn:

> "You need to know what you are talking about."

DFT mission: understand what the core of the data domain is, develop definitions of core terms based on useful data models.

DFT is part of coming to an agreed upon culture in RDA.

Scope:

We only speak about the **domain of registered data**.

knowing that there is a lot of non-registered data

knowing that some disciplines are necessarily further away from what we are discussing

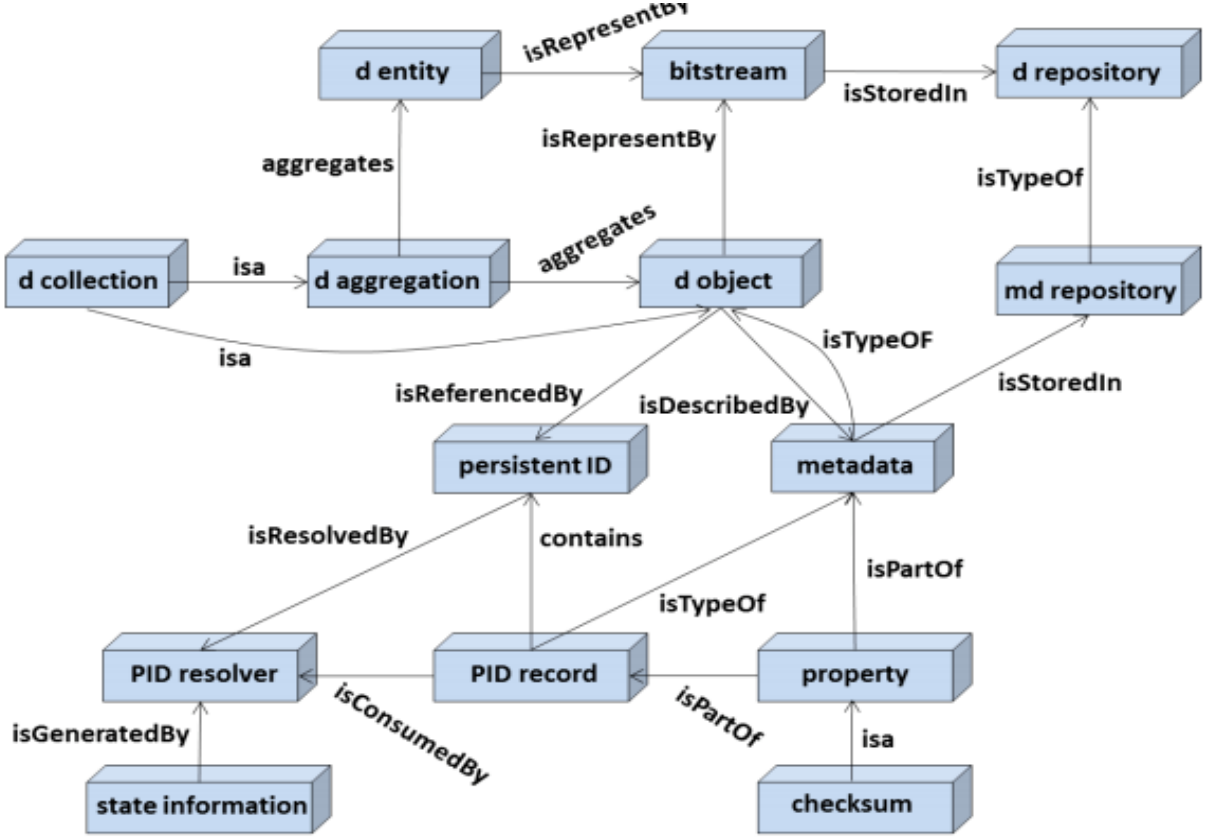
DFT: The Original Problem

Data Management & Processing remains time consuming & costly due in part to the **heterogeneity of data organized** in particular with respect to **logical information** and how it is **documented**.

Researchers see the need to change habits & routines, but do not have **agreed on common models and associated terminology, policy and best practices** that can be used across communities and stakeholders to discuss data organization, sharing and re-use.

Our goal was to define data management terms clearly enough to help mitigate the communication challenges.

DFT: Synthesis Model



DFT: Take-home Message

Even if you don't care about all the theory: If there is one thing to highlight from all that activity that is

Register your data

Aka, get it a PID

Why not just a URL?

You want the identifier to

- Be independent of
  - network location
  - domain owner
- Hold some metadata on, e.g.
  - replicas or other versions
  - where to find metadata
  - checksum
  - data type …

The Persistent Identifier Type Registry

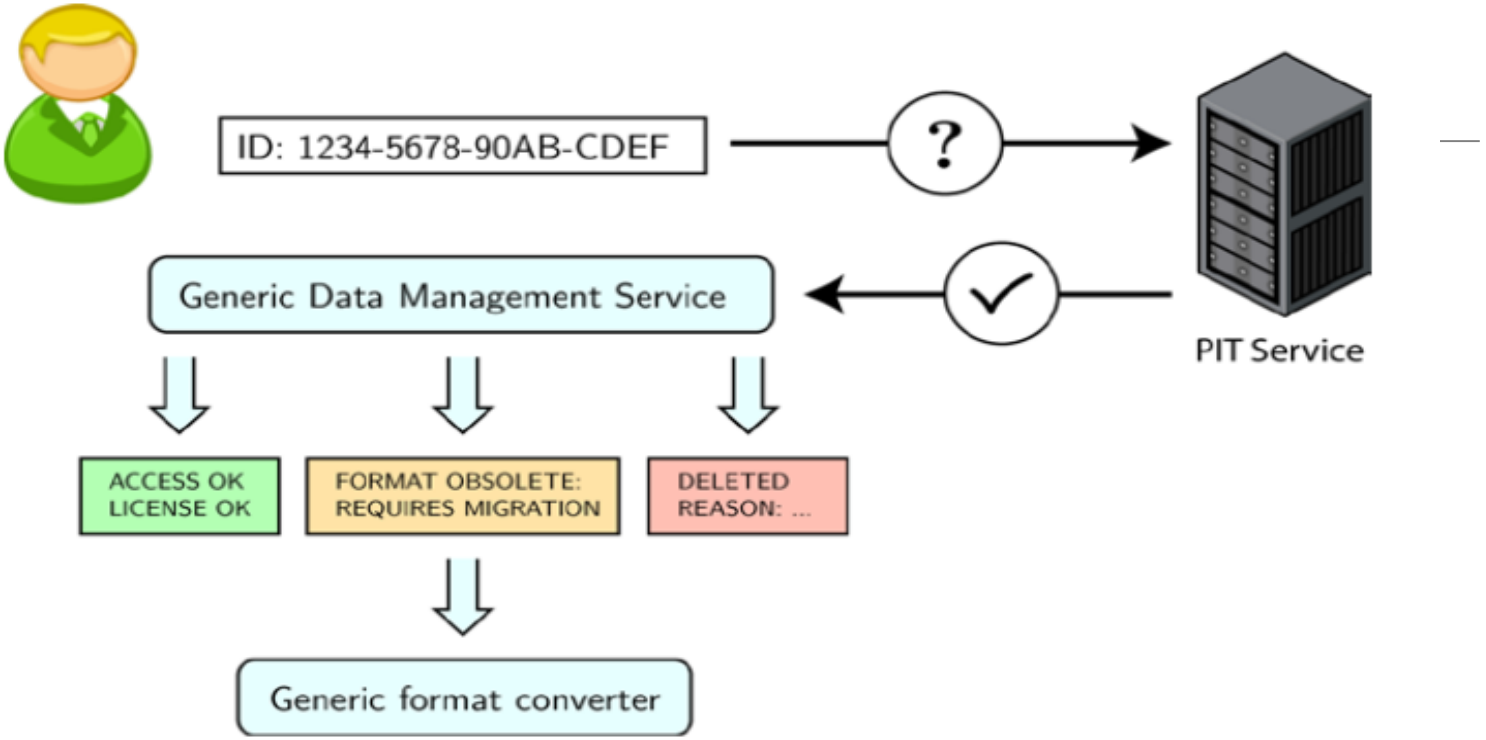


Figure 1: An exemplary workflow initiated by an end-user encountering an identifier to an unknown resource. Using a generic service endpoint designed for such end-user queries, the identifier will be resolved and typed information returned to an intermediate data management service. This service can then decide upon the typed information on the status of the identified resource and finally redirect the user to services matching both the resource type and its current status.

Data Type Registries (DTR): Introduction

Understanding scientific data and metadata is hard

- **Researcher 1:** "Could you tell me what **column 12** means in the CSV file you referenced in paper A **from 5 years ago**?"
- **Researcher 2:** "Uh, I **believe** it's a **number**"
- **R1:** "I can see that. Could it be **a temperature**?"
- **R2:** "Probably"
- **R1:** "**Fahrenheit**? **Celsius**?"
- **R2:** "Maybe **Kelvin** or **Rankine**?"
- **R1:** "Kelvin?"
- **R2:** "On second thought, maybe **it's not** really **a temperature**"
- **R1:** "…"

DTR: The Problem

Automatically analyzing and processing scientific data and metadata is even harder

- What is sequence **"00010101010000100101111 10"**?
- It could be an ***integer*** → how many bits?
- It could be a ***floating point number*** → precision?
- It could be a ***string*** → encoding?
- **Even if we knew: What does it represent?**

What is a Data Type Registry?

A DTR is a low-level service/infrastructure with the ability to record and disseminate "Data Type Records"

What is a Data Type Registry?

A DTR is a low-level service/infrastructure with the ability to record and disseminate "Data Type Records"

But, What is a Data Type?

What is a Data Type?

A Data Type is a characterization of data at any level of granularity

- From small individual observations to large structured datasets
- can include (aka make reference to) other data types

Must include information about structural organization, contexts and assumptions in the data

- Cell A3 is a number, but is it a temperature? Celsius?
- It's a dataset, but what are the variable names?
- Is it packed as CSV/NetCDF? A single unit? A collection?

Must be permanently linked to the described data

Should be standardized, unique and discoverable

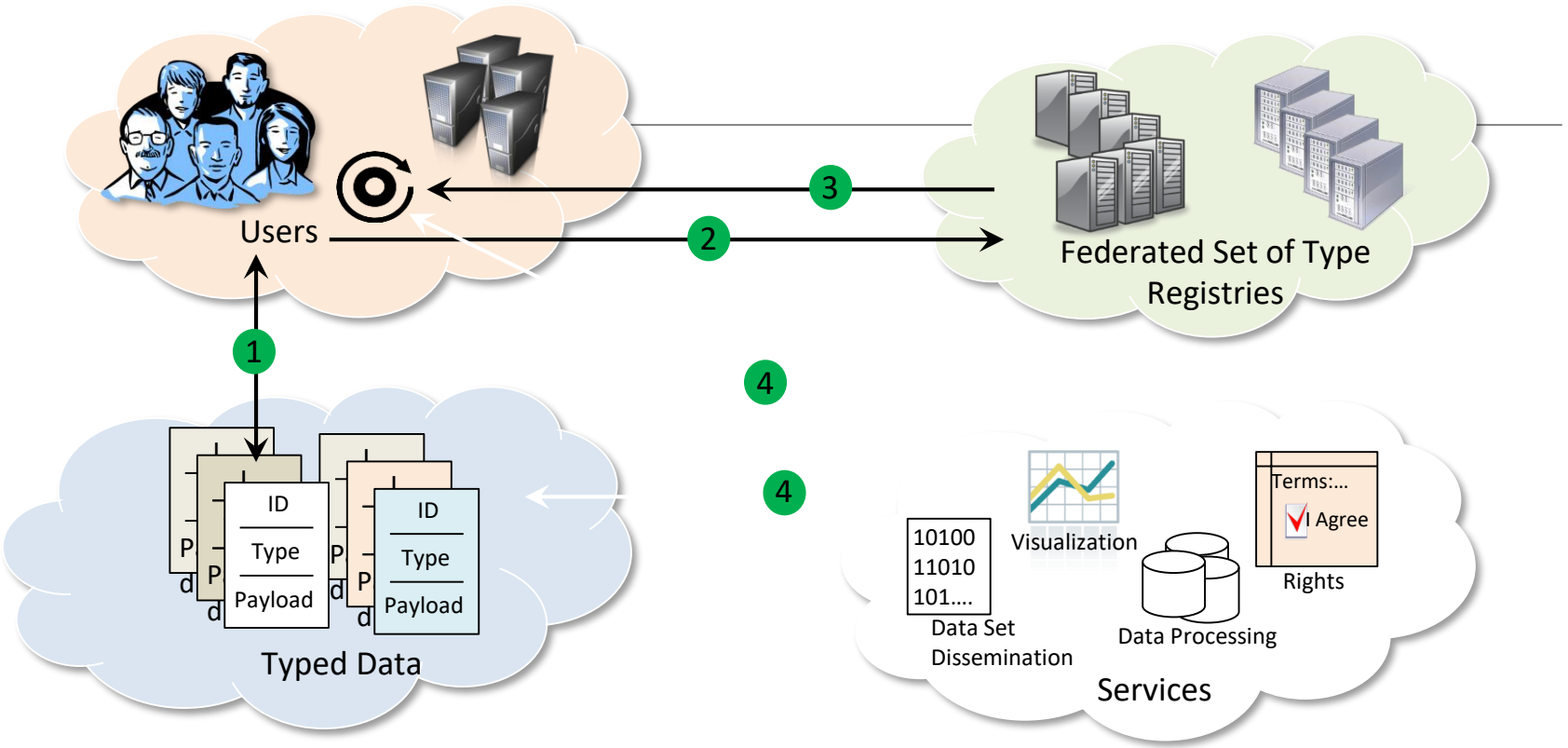
What is a Data Type Registry?

A DTR is a low-level service/infrastructure with the ability to record and disseminate "Data Type Records"

Minimum requirements:

- Should assign **unique and resolvable identifiers** to created/stored Data Type records
- Should **enforce and validate** a common **data model** for describing Data Types and their structure
- Should **allow interoperability** between multiple instances
- Should offer a **UI for human use**
- Should offer an **API for machine use**

DTR Examples: Processing Use Case



1. Clients (processes or people) encounter an unknown type
2. The Type is resolved to the Data Type Registry
3. Response includes type definitions, relationships, properties, and possibly service pointers. Response can be used locally for processing, or, optionally …
4. Typed data or references to typed data can be sent to service provider

Source: "Data Types" Giridhar Manepalli, RDA 2nd Plenary

DTR: Quick Links

- At CNRI - lead developer: http://typeregistry.org/
- Cordra - a generic registry application that can be used to easily implement a DTR: https://cordra.org/
- RDA WG https://www.rd-alliance.org/groups/data-type-registries-wg.html
- At ePIC: http://dtr.pidconsortium.eu/
  - Test instance: http://dtr-test.pidconsortium.eu/

RDA Practical Policy Working Group Focus

Identify the most important policies

Practical implementations for managing research data collections

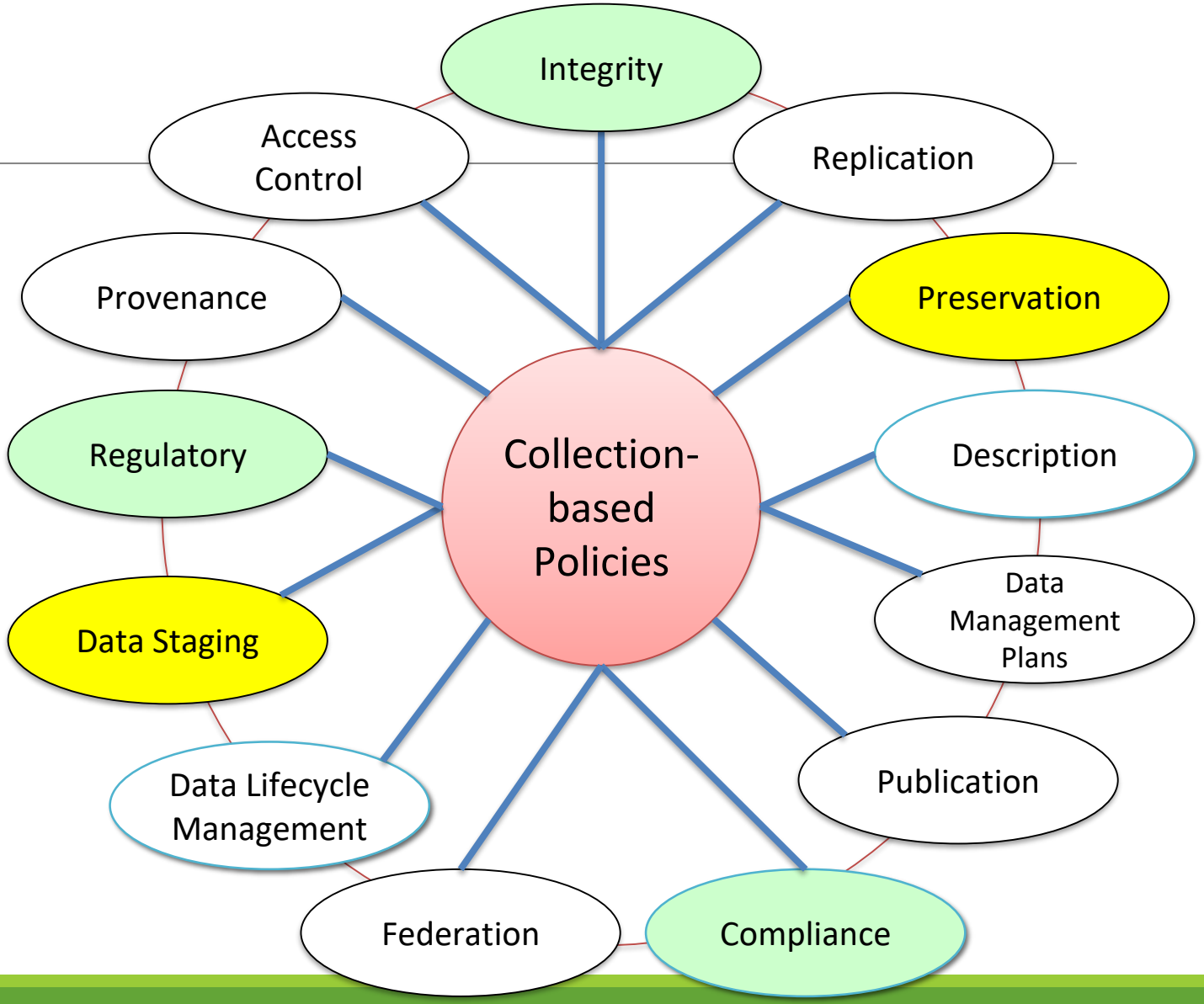
Provide recommendations for a “starter kit”

Testbeds:

- Evaluate standard policies
- Test interoperability across WGs

Policy:
Assertion or assurance that is enforced about a collection or a dataset

Management
Administrative
Assessment



Summary of policies in production use

1. Policy for data retention. How long, how short? Need preservation, or not? (5) Retention and disposition
2. Notification policies. (Ex. must warn data researcher that their data will be deleted at X time.) (6) notification on event
3. Transferability policies. The data must be transferable from the repository back to the researcher and the repository of origin. Or, in the event of defunding, the data must be de-accessioned and moved to another repository (or not, depending on relevant SOPs, agreements, etc.).
4. Policies re: costs and who pays for all of this data storage (8)
5. Policies around context. Sometimes the original data and additional metadata are needed. Sometimes, the context or derived data is what matters, and not the data itself. (7)

Summary of policies in production use (cont)

6. Policies re: tagging/annotating data
7. Search/Information Retrieval policies. What parts of the data will you search on, or not search on? (4) Controlling search
8. Standard Sys Admin policies: (1) replication, back up, (2) integrity checks, syncing with back ups.
9. Content policies: do we care what content and file formats users upload? Some do, some don't. (3) Transformative migration
10. Policy to educate researchers about all of the different policies relevant to the data repository. For example, a user agreement/Terms & conditions statement that researchers must check off.

Best Practices for production policies

Consensus on a policy

- Use at multiple institutions
- Generality

Best practice policy components

- Name of operation that policy controls
- Constraints that policy implements
- State information that policy uses or modifies
- Verification policy
- Example of running code
- Documentation

Operations managed by policies

Paper posted that lists 70 operations

- Policy-verification.docx

Candidate operations

- Access control
- Backups
- Data retention
- Descriptive metadata
- Format creation
- Integrity checks
- Notification
- Policy constraints
- Replication
- Restricted search
- Storage cost
- Tags
- Use agreements

Policy Types (exp)

Policy type	Operation
Access	Set access control
	Check access control
	Audit access control
Backups (time-stamped copies)	Create copy
	Set timestamp
	Verify timestamps
Contextual metadata	Extract metadata
	Register metadata
	Verify metadata
Data Retention	Set retention period
	Check retention
	Verify retention
Disposition	Define migration location
	Migrate data
	Verify migration

Practical Policy WG Results

- Summary of results
- Policy templates
- Policy verifications
- Policy implementations (mostly iRODS, some GPFS)

Available from

https://www.rd-alliance.org/filedepot?cid=104&fid=553

Contacts: Reagan Moore and Rainer Stotzka

Next 5 Technical Specifications under evaluation

EVALUATION MSP MEETING
28 SEPT 2017

✓TS5: **Dynamic-data Citation Methodology** Supports efficient processing of data and linking from publications.

✓TS6: **Data Description Registry Interoperability Model**: Interoperability model addressing the problem of cross platform discovery by connecting datasets together.

✓TS7: **RDA/WDS Repository Audit and Certification Catalogues**: Creates harmonized Common Procedures for certification of repositories at the basic level, drawing from the procedures already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSU-WDS)

✓TS8: **RDA/WDS Workflows for Research Data Publishing Model**: A data-publishing reference model assisting research communities in understanding options for data publishing workflows and increases awareness of emerging standards and best practices.

✓TS9: **RDA/WDS Publishing Data Services**: An open, universal literature-data cross-linking service to improve data visibility, discoverability, re-use and reproducibility

Next 5 Technical Specifications under evaluation

EVALUATION MSP MEETING
28 SEPT 2017

✓**TS5: Dynamic-data Citation Methodology Supports efficient processing of data and linking from publications.**

✓TS6: **Data Description Registry Interoperability Model**: Interoperability model addressing the problem of cross platform discovery by connecting datasets together.

✓TS7: **RDA/WDS Repository Audit and Certification Catalogues**: Creates harmonized Common Procedures for certification of repositories at the basic level, drawing from the procedures already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSU-WDS)

✓TS8: **RDA/WDS Workflows for Research Data Publishing Model**: A data-publishing reference model assisting research communities in understanding options for data publishing workflows and increases awareness of emerging standards and best practices.

✓*TS9: **RDA/WDS Publishing Data Services**: An open, universal literature-data cross-linking service to improve data visibility, discoverability, re-use and reproducibility => postponed*

From Data to Knowledge

Call for Poster Sessions https://www.rd-alliance.org/rda-11th-plenary-poster-session

ends 1st February, midnight UTC

To find out more visit:

https://www.rd-alliance.org/plenaries/rda-eleventh-plenary-meeting-berlin-germany

RDA in a Nutshell

WWW.RD-ALLIANCE.ORG/
@RESDATALL

RDA Global

Email - enquiries@rd-alliance.org

Web - www.rd-alliance.org

Twitter - @resdatall

LinkedIn - www.linkedin.com/in/ResearchDataAlliance

Slideshare - http://www.slideshare.net/ResearchDataAlliance

RDA Europe

Email - info@europe.rd-alliance.org

Twitter - @RDA_Europe

RDA US

Twitter - @RDA_US