Digital Objects and the RDA Data Fabric Approach

ITU Workshop and fifth FG-DPM meeting

Peter Wittenburg (MPCDF), Ulrich Schwardmann (GWDG)

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

Am Fassberg, 37077 Göttingen
ulrich.schwardmann [at] gwdg.de

17 September 2018, Tunis

Content

research data sharing without barriers
rd-alliance.org

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach

Digital Objects

FAIR

Findability and
PID
Reusability and
Metadata
Interoperability
and Registration
of Types

Accessibility &
the DO Cloud

DOIP
Collections
DO Browser

Trust   2 / 31

Dynamics in the Data Domain

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach
Digital Objects

FAIR
Findability and
PID
Reusability and
Metadata
Interoperability
and Registration
of Types

Accessibility &
the DO Cloud
Collections
DO Browser

Trust   3 / 31

- Data heterogeneity hampers data exchange and reuse already now.

- about 80% of the time of data experts is wasted with data wrangling (i.e. making data ready for analytics),
  - RDA EU 2013 Survey: 75%
  - M. Brodie MIT S.: 80%
  - CrowdFlower 2017 S.: 79%
  - findings of relevant data analytics projects:
    - cancer, climate, environment, physics, humanities, etc.

- In industry the phenomena are essentially the same
  - BD/AI Summit 2018: 60% of industrial data projects fail

- All will become even worse with IoT and new sensors

Abstractions in the Data Domain

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

- the mayor obstacle for automation:
  Heterogeneity and complexity of data
- Abstraction
  - is a way to hide heterogeneity and complexity
- Virtualisation
  - provides a layer of abstraction between data and application
  - in our case the reference becomes a placeholder for data
- Encapsulation
  - provides a layer of abstraction between inner heterogeneity and complexity and outer simplification
  - in our case the reference becomes the broker for information about inner complexity

Digital Objects

The Core Data Model of the
RDA Data Foundation & Technology Working Group
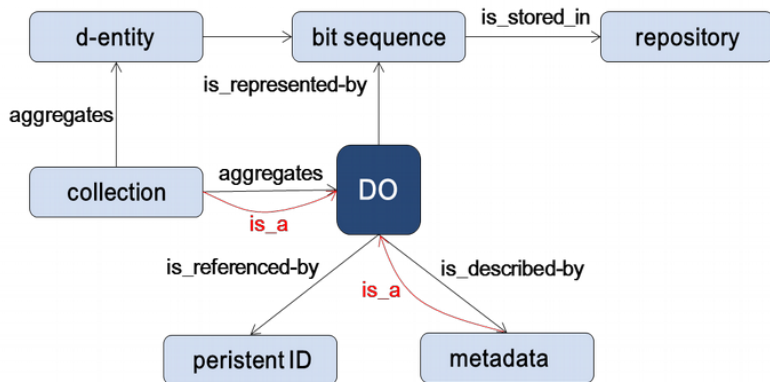
Digital Objects and Persistent Identifiers Virtualisation

The central role of the PID reference.
A similarity to the IP in the network technology
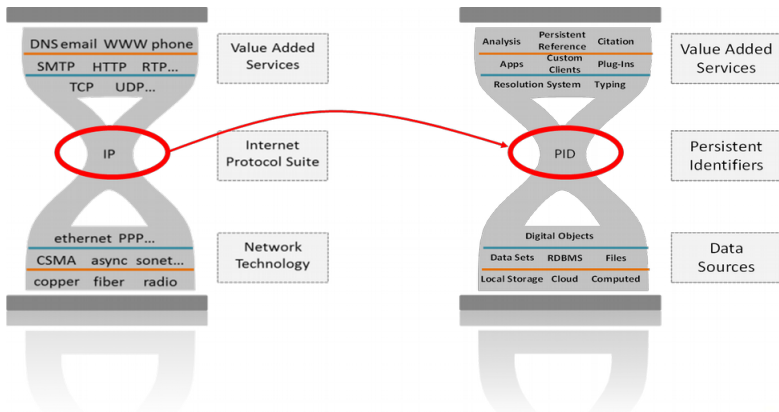
Digital Objects and Persistent Identifiers
Encapsulation

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

The RDA PID Bindings

research data sharing without barriers
rd-alliance.org

The FAIR Data Guiding Principles

- **F**indability
- **A**ccessabily
- **I**nteroperability
- **R**eusability

see also:

https://www.force11.org/group/fairgroup/fairprinciples

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

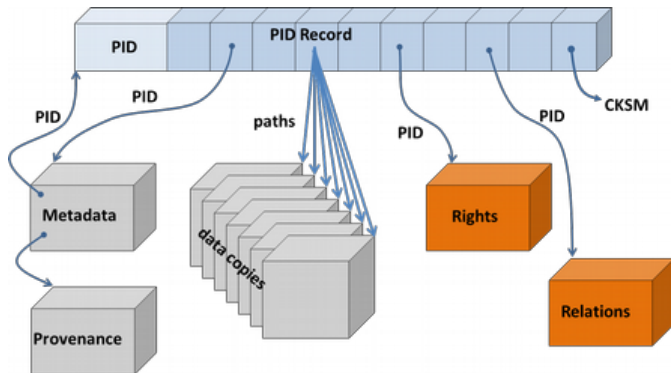The RDA
Data Fabric
Approach

Digital Objects

FAIR

Findability and
PID

Reusability and
Metadata

Interoperability
and Registration
of Types

Accessibility &
the DO Cloud

DOIP

Collections

DO Browser

Trust   8 / 31

Findability
Persistent Identification and Redirection

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

- URLs and cool URLs turned out to be highly instable
- PURL: persistent URLs, based on HTTP-redirection
  - central solution or administration/ownership unsolved
  - not reliable anymore by **organisational instability**
- better use **redirection** provided by a distributed system
- examples: URN, ARK, **Handle** (incl. DOI)

Metadata Description → Handle → Handle Service → URL → Electronic Resource
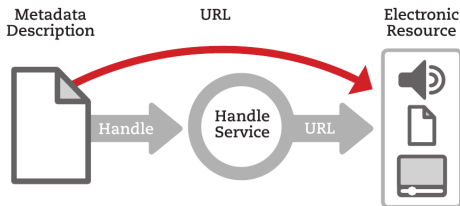
Handle-PIDs:
21.11234/12345678
[Prefix]/[Suffix]

- **Critical:** Redirection always needs Resolution

research data sharing without barriers
rd-alliance.org

Reusability

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

- needs knowledge about basic properties of data
  - **Metadata** is often unavailable, not connected to data or not interpretable
- For reuse provide as much of this knowledge before access to the data
  - Data Format Migration needs information about the format
- Registration:
  bind metadata and data with PID to a digital object

PID Information Types

- are additional metadata, stored in the PID database
- intended to be directly accessible without any redirection
- similar to mime types, typical examples are:
  - checksum
  - mime type
  - reference information
  - versioning (relative and absolute)
  - embargo time
  - expiration date
  - add. metadata location
  - basic Dublin Core
  - access restrictions and methods
  - data and table column formats
  - collection description
  - ...
- there will be more and others for IoT

Data Fabric
and DOs
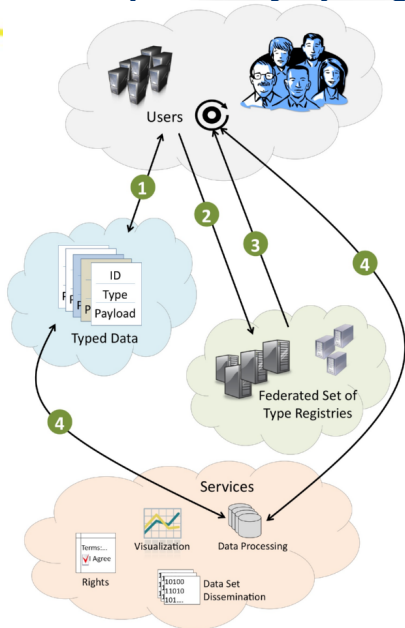
Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach
Digital Objects

FAIR
Findability and
PID
**Reusability and
Metadata**
Interoperability
and Registration
of Types

Accessibility &
the DO Cloud
DOIP
Collections
DO Browser

Trust   11 / 31

Interoperability by Registration of Types

RDA working group on
Data Type Registries

- approach to provide *type definitions*
- a PID for each definition
- defines the type structure, its use and semantics
- CORDRA as DTR service
- typical use cases:
  - with given PID find a type and ask for its use at DTR (see left)
  - ask at DTR for types with given semantics and find via PIDs according data

The ePIC Data Type Registry

- **Features**
  - Definition of PID Information Types
  - hierarchical types and automated schema extraction
  - Access via REST API, Browser
- based on CORDRA software
- GWDG is provider on behalf of ePIC
- Who can use the service?
  - public, authorization needed only for type definition

Overview: `http://dtr.pidconsortium.eu/`

PID InfoType states are:

- *in preparation* (21.T11148),
  - `http://dtr-test.pidconsortium.eu/`
- *candidate, approved, deprecated* (21.11104)
  - `http://dtr-pit.pidconsortium.eu/`

ePIC Data Type Registry   Introduction   All   Types ▾   Sign In

ePIC
Persistent Identifiers for eResearch

GWDG

Search   Search

rd-alliance.org

hierarchical type definitions

- types are often dependent from each other, how exactly?
- to exactly describe JSON objects by data types one needs:
  - a distinction between derived objects and basic objects
    - concept of *basic PID info types* and *PID info types*
  - a more exact description of the type dependencies
  - additionally a JSON schema inspired dependency model
- in consequence:
  - possibility to derive JSON schemas for the type values
    - automated server side schema derivation at ePIC DTR
  - one type defines in an exact way its whole dependencies
    - in objects of a certain type one can use the names of its parts (instead of type identifiers)
- see also Schwardmann, U.: Automated schema extraction for PID information types
  - PID: http://hdl.handle.net/21.11101/0000-0002-A987-7

Accessibility and the DO Cloud

Accessibility has technical and social aspects,
Digital Objects can help in both aspects

- technical
  - access rights are a kind of metadata, PID points to it
  - access rights can be made explicit as PID types
    - at least the authorization part, like read/write rights

- social
  - PID can help to make access **more transparent**
    - embargo as type
    - explicit access rights and methods
  - which rights are granted for research data is a **scientific decision**

The Digital Object Cloud

The users view of the DO Cloud

End users, developers, and automated processes

deal with persistently identified, consistently structured digital objects

which are securely & redundantly managed & stored in the Digital Object Cloud

which is an overlay on existing or future information storage systems.

Accessibility &
the DO Cloud
DOIP
Collections
DO Browser
Trust   16 / 31

---

[1] Global Digital Object Cloud, Larry Lannom, 2016

rd-alliance.org

The Digital Object Cloud

The users view of the DO Cloud

| End users, developers, and automated processes | deal with persistently identified, consistently structured digital objects | which are securely & redundantly managed & stored in the Digital Object Cloud | which is an overlay on existing or future information storage systems. |

Data Fabric and DOs

Accessibility & the DO Cloud
DOIP
Collections
DO Browser
Trust 17 / 31

---

[2] Global Digital Object Cloud, Larry Lannom, 2016

The Digital Object Interface Protocoll **DOIP**
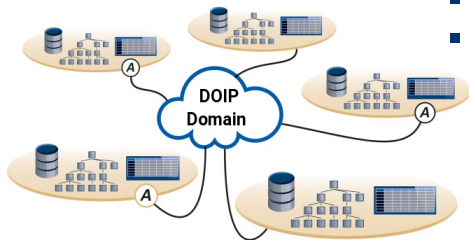
- defined in ITU x.1255
- describes
  - the DO-cloud with
    - digital objects
    - provided by repositories,
    - referenced by PIDs
  - with metadata types,
  - registered in DTRs,
  - **and registered services**
    - again referenced by PIDs

Collections

- Collections structure the access to components

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

What are Collections in the RDA sense?

- Abstractly they are PIDs pointing to a list of PIDs
  - and additional metadata to enable services
  - this is a **recursive** definition: members can be collections
- the RDA outcome is a concrete REST API to manage collections
- collections are ubiquous also in data management:
  - directories, zip and tar archives, ...
  - objects structured by chapters, pages, newlines, ...
  - group definitions, ...
- collections are a very general way to organize objects hierarchically
  - PIDs are a completely flat view on global objects
  - the RDA collection helps to build hierarchies on objects
  - they only need **names as additional metadata** to make sense also for humans
- often repositories have an implicit hierarchical structure

A Collection Repository

A Collection Repository

Handle.Net®

Handle Values for: 21.T11998/0000-0002-859F-5

Index	Type	Timestamp	Data
1	URL	2017-09-12 11:04:09Z	http://coll-reg.pidconsortium.net/21.T11998/public/collections/21.T11998/0000-0002-859F-5
2	21.T11148/3ed964ae3f96d067f2ac	2017-09-12 11:04:09Z	{"properties": {"description": "3rd order collection"}, "capabilities": {"isOrdered": false, "supportsRoles": false}}
3	21.T11148/ec9db37ca4b137579592	2017-09-14 16:25:02Z	[{"id": "21.T11998/0000-0002-859C-8", "mappings": {"index": 2}}, {"id": "21.T11998/0000-0002-859E-6"}]
100	HS_ADMIN	2017-09-12 11:04:09Z	handle=21.PERS/0000-0011-2D9D-B; index=300; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]

research data sharing without barriers
rd-alliance.org

A Collection Repository

- is **repository agent** in the Digital Object Access Protocol
  - maintains a repository based on type entries in the collection PID
  - defines adaptor classes for different collection like structures

- is **adaptor** in the Digital Object Access Protocol
  - DARIAH repository (humanities)
    - presents collections based on PIDs, but has no RDA collection API
  - IPCC-EFDB emmission factor repository (climate research)
    - collection PIDs provided by ePIC Collection Repository
    - endpoints provided by ePIC PID service
  - ITIS taxonomy (biology) (in preparation)
    - based on unique and stable internal reference numbers
    - implementation via templates or fragment identifiers

A Digital Objectes Browser

Trust, Reliability and Sustainability

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach
Digital Objects

FAIR
Findability and
PID
Reusability and
Metadata
Interoperability
and Registration
of Types

Accessibility &
the DO Cloud
DOIP
Collections
DO Browser

- Quality of digital repositories: **CoreTrustSeal**
- Focus now on trust of mission critical registries
  - PID/Handle Resolution
    - Policies and QoS in ePIC
  - Data Type Registries
    - Policies for Data Type Life Cycle
- Other mission critical registries for example
  - registries storing access permissions,
  - license signatures, etc.
- Handle Resolution
- Policies and QoS

Digital Object Numbering Authority
DONA

Data Fabric and DOs

Peter Wittenburg (MPCDF), Ulrich Schwardmann (GWDG)

Dynamics in the Data Domain

The RDA Data Fabric Approach

Digital Objects

FAIR

Findability and PID

Reusability and Metadata

Interoperability and Registration of Types

Accessibility & the DO Cloud

DOIP

- DONA governs the Handle structure at the top level in close collaboration with ITU
  - the Multi Primary Administrators (MPAs) build the distributed Global Handle Registry (GHR)
- DONA is a Swiss foundation hosting an international consortium
  - DONA was founded 2012 in Geneva
  - currently 8 credentialed MPAs
  - ITU is involved as partner
  - GWDG is MPA on behalf of ePIC
  - The DONA consortium will moderately grow in future

research data sharing without barriers
rd-alliance.org

RDA

RESEARCH DATA ALLIANCE

Collections

DO Browser

Trust   26 / 31

The Handle Resolution System

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

The **ePIC** Persistent Identifier Consortium for eResearch

is a network of currently eight strong scientific service providers that signed a contract,

Data Fabric and DOs

Peter Wittenburg (MPCDF), Ulrich Schwardmann (GWDG)

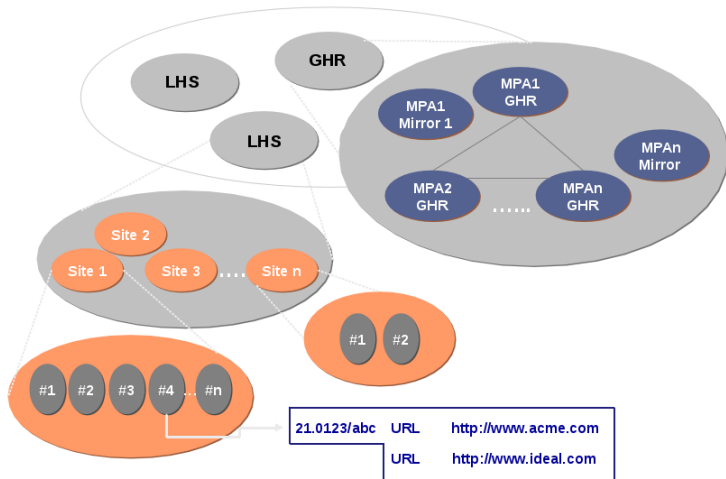Dynamics in the Data Domain

The RDA Data Fabric Approach

Digital Objects

FAIR

Findability and PID

Reusability and Metadata

Interoperability and Registration of Types

Accessibility & the DO Cloud

DOIP

Collections

DO Browser

Trust

ePIC
Persistent Identifiers for eResearch

CSC

CSCS

Computer Network Information Center
Chinese Academy of Sciences

DKRZ

grnet

GWDG

SNIC
research data sharing without barriers
rd-alliance.org

SURF SARA

- to **ensure a reliable and persistent identifier infrastructure**,
- devoted to the needs of the research community at large.
- Quality of Service
- **Mayor focus**: referability
  - for sharing data during the research process
  - with finer granularity and
  - PID coupled metadata

Policies and Quality of Service in ePIC

- Conditions of Operation
  - user management, privacy protection and secrecy
- incident management and monitoring
- support system with agreed responsabilties
- certification of ePIC PID services
- several policies for PID minting and update agreed
  - others are still under discussion
- quality of resolution
  - audits can be requested
- community dependend policies (on prefix level)
- the ePIC members provide PID services for their customers

  - implementation of an ePIC Data Type Registry
  - implementation of an ePIC Collection Repository

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach

Digital Objects

FAIR

Findability and
PID

Reusability and
Metadata

Interoperability
and Registration
of Types

Accessibility &
the DO Cloud

DOIP

Collections

DO Browser

Trust 29 / 31

RDA

RESEARCH DATA ALLIANCE

Policies for the ePIC DTR Type Life Cycle

- **in preparation** prerequisites:
  - for types in preparation
    - should use *(basic) PID info types* and a schema
- **candidate** prerequisites:
  - provenance
    - an existing *in preparation* type
  - consistency
    - need: name, description
    - special reqirements for PID-InfoTypes and PID-BasicInfoTypes
    - governance: an application to become a candidate is needed
- **approved** prerequisites:
  - provenance
    - needs a candidate
- **deprecated** prerequisites:
  - provenance
    - needs an approved type, a reason for replacement
    - a mutual reference inside the new and the old type

Many Thanks

Data Fabric
and DOs

Peter
Wittenburg
(MPCDF),
Ulrich
Schwardmann
(GWDG)

Questions ???

Dynamics in
the Data
Domain

The RDA
Data Fabric
Approach

Digital Objects

FAIR

Findability and
PID

Reusability and
Metadata

Interoperability
and Registration
of Types

Accessibility &
the DO Cloud

DOIP

Contact at ePIC:

- support [at] pidconsortium.eu

Contact at GWDG:

- **Ulrich Schwardmann**
  T: 0551 201-1542, E: ulrich.schwardmann [at] gwdg.de

research data sharing without barriers
rd-alliance.org