



# Data Type Registries #2

Larry Lannom / Tobias Weigel  
22 March 2018

## RDA P11: Recommendations and Outputs Session

---

# Summary of the Problem

---

- Data sharing requires that data can be parsed, understood, and reused by people and applications other than those that created the data
- How do we do this now?
  - For documents – formats are enough, e.g., PDF, and then the document explains itself to humans
  - This doesn't work well with data – numbers are not self-explanatory
    - What does the number 7 mean in cell B27?
- Data producers may not have explicitly specified certain details in the data: measurement units, coordinate systems, variable names, etc.
- Need a way to precisely characterize those assumptions such that they can be identified by humans and machines that were not closely involved in its creation
- Affects all data producers and consumers

- 
- See output of DTR #1 – adopted as ICT Tech Spec
    - Confirmation that detailed and precise data typing is a key consideration in data sharing and reuse and that a federated registry system for such types is highly desirable and needs to accommodate each community's own requirements
    - Deployment of a prototype registry implementing one potential data model, against which various use cases can be tested
    - Uptake in a number of projects and other RDA groups
  - Current State
    - Continued spread of the fundamental concept
    - Multiple type registries in (mostly experimental) operation
    - No agreed-upon standard minimal type record
    - No federation across registries or governance structures
    - All of this to be discussed in DTR Breakout later today.
    - Draft final report on DTR #2 still in motion
    - ISO interest – more later

# Impact of the Recommendation

---

- Data Typing is Infrastructure for Infrastructure
- Impact will vary by discipline/process
- Examples
  - Types of returned PID resolution values, e.g., checksums, public keys, to inform subsequent actions
  - Precise characterization of data sets for interpretation and re-use
  - Precise characterization of entities in a workflow to determine routing to next appropriate stage
  - Object type as a search criteria
  - Development and advertising of services relevant to given sets of object types
- Open availability of types via registries could encourage greater standardization across data sets in order to reuse types and associated services and software

# Endorsements/ Adopters

---

- ICT Technical Specification
- Multiple RDA Groups and Affiliated Projects
  - Data Fabric IG
  - PID Kernel Information WG
  - Collections WG
  - RPID (NSF project at IU, Tufts, CNRI)
  - DCO
  - C2CAMP
  - Vermont Monitoring Cooperative
  - ePIC / GWDG
  - Others.....
- ISO SC32 WG2
  - Is DTR new or should it be part of ISO 11179?
  - RDA needs to help drive this – Denise Warzel will explain