PIT services: the what and the why

Beth Plale, Jason Haga, Quan (Gabriel) Zhou

9/8/16

RESEARCH DATA ALLIANCE

First: Building a concrete data fabric configuration

Beth Plale, Tobias Weigel

Taken from RDA PID Training, Garching, 2016/08/31

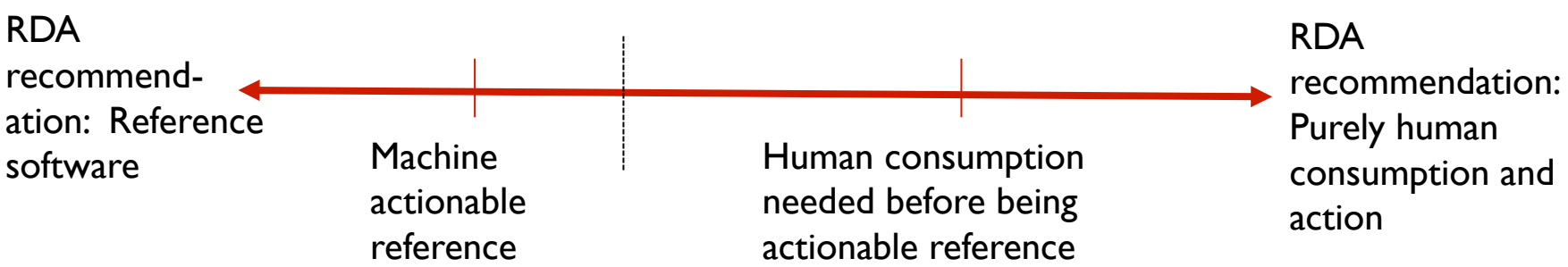
research data sharing without barriers

rd-alliance.org

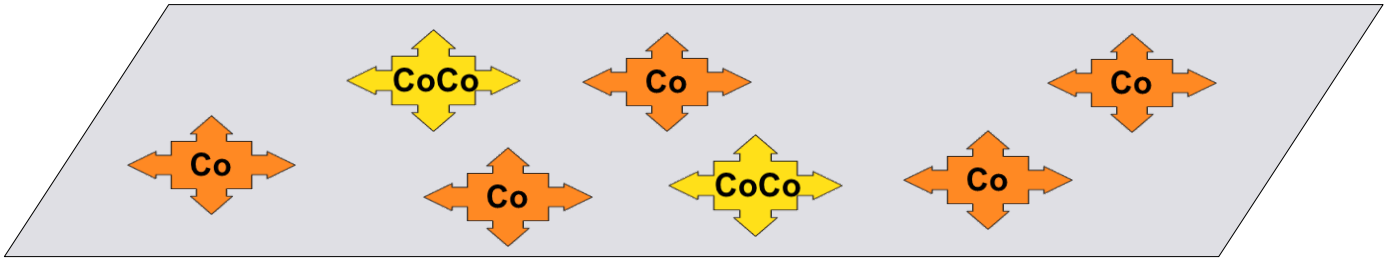
- RDA Data Fabric has activity that examines fabric composition
  - Composing from RDA Recommendations (largely but not exclusively)
- A couple Recommendations are around PIDs (session view not citation view)
- Inductive (direct) approach to composition of component

Dimensions of Testing: getting to Data Fabrics

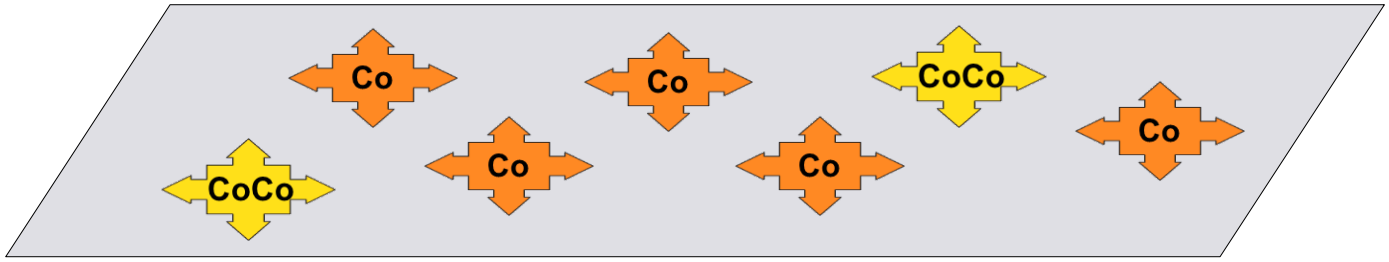
- RDA produces RDA Recommendations (i.e., outputs)
- Some technically-oriented RDA Recommendations have reference software with it, these are starting point
- Many technically oriented RDA recommendations do not have reference software, yet are machine actionable (a schema for instance). These also are of immediate interest to data fabric composition.
- Other recommendations play important but background roles in early composition



Compositions of components

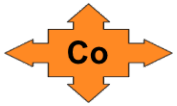


Composition (or Fabric) A



Composition B

Core Components & Services



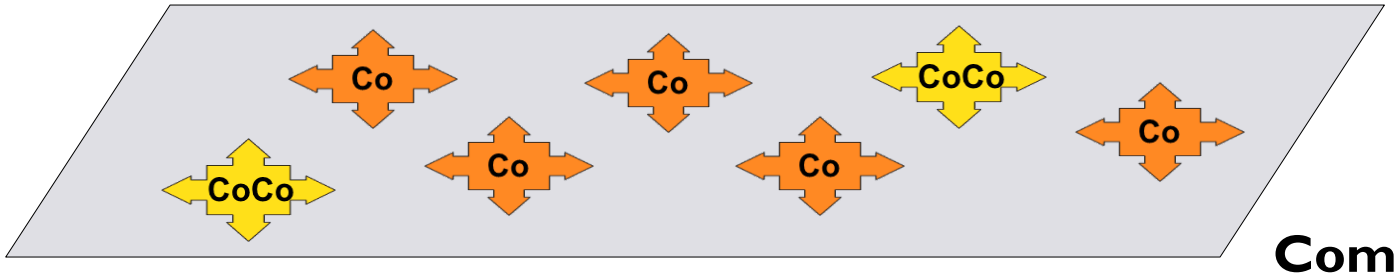
Specific Components & Services

Compositions of components

Given nature of data (can't do much without understanding it), successful data fabric will likely:

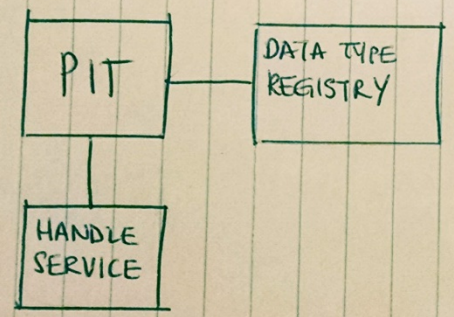
1. Run on possibly distributed e-infrastructure (EUDAT, NDS, …)
2. Serve scholarly domain as domain infrastructure
3. Support multiple projects within that domain
4. And eventually result in cross-domain research

For 3 and 4 to be realized, composition of 2 must be shared across projects

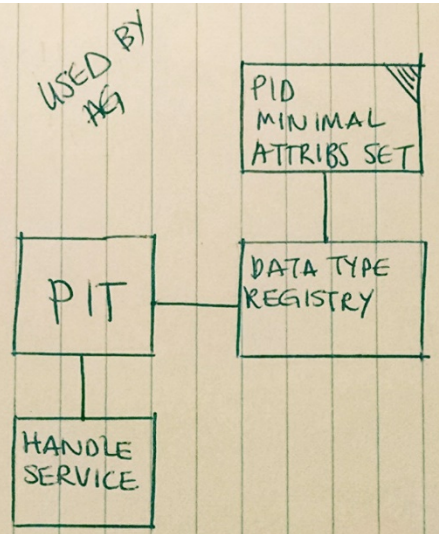


Composition (Fabric) B

- RDA PIT WG Recommendation and RDA Data Type Registry Recommendation are starting point for composition, hence my interest in the topic

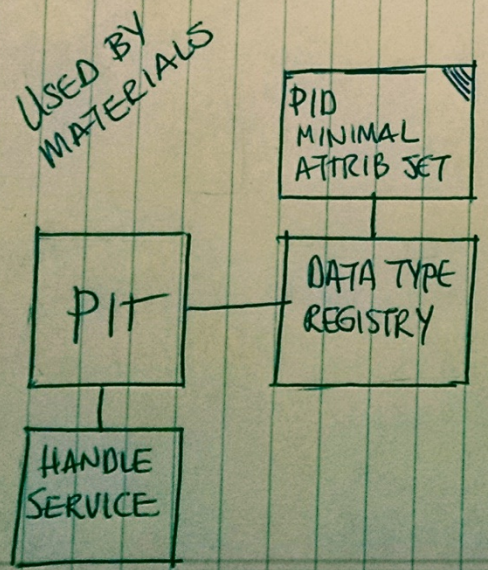


COMPOSITION (FABRIC) 1

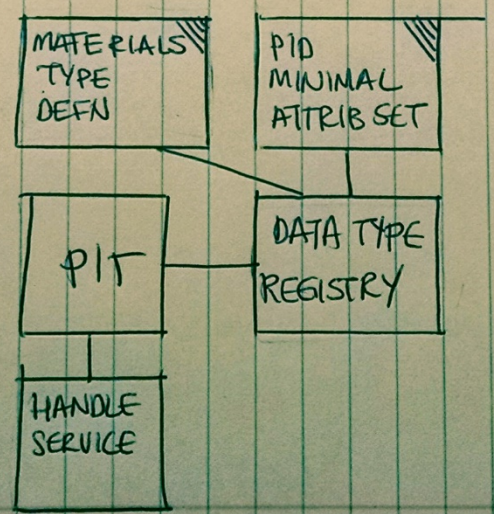


COMPOSITION 2

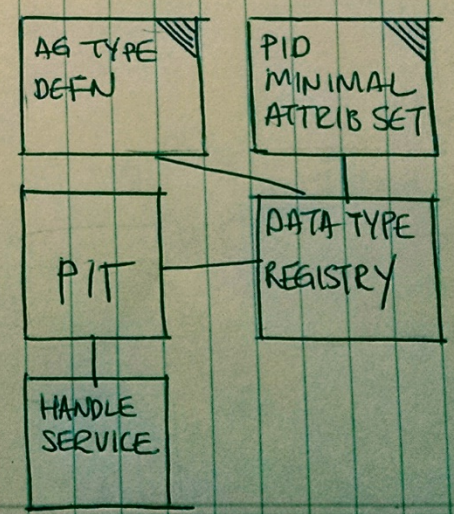
- COMPOSITION 1 IS NOT USEABLE AS IS
- COMPOSITION 2 ≡ COMPOSITION 3 SO SUBJECT TO "CORE" DESIGNATION
- COMPOSITION 4 AND COMPOSITION 5 START TO GET AT UNIQUE COMMUNITY INSTANCES OF A FABRIC
- ▨ : DOESN'T EXIST TODAY
- COMPOSITIONS 4 AND 5 HAVE N=5



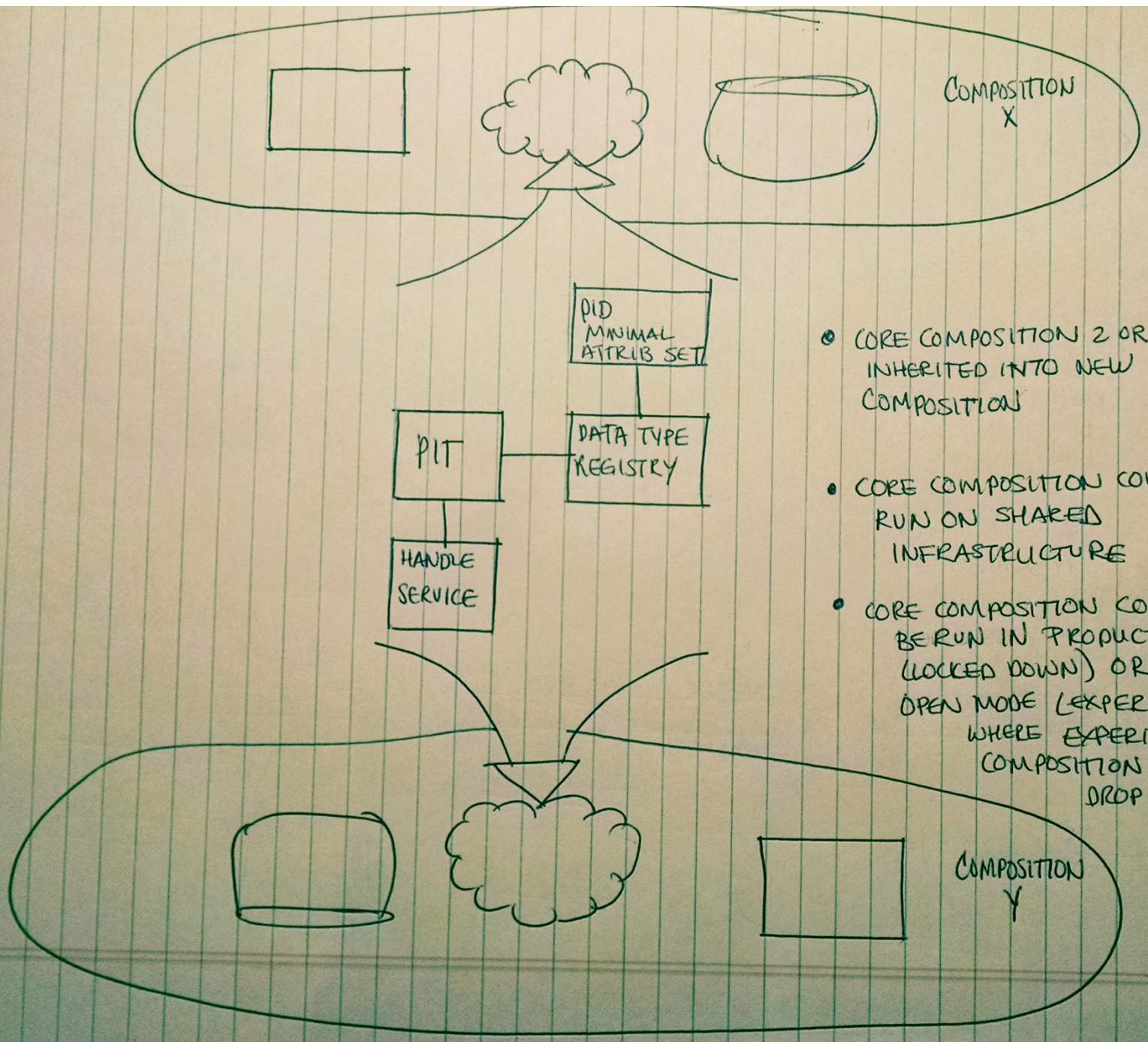
COMPOSITION 3



COMPOSITION 4



COMPOSITION 5

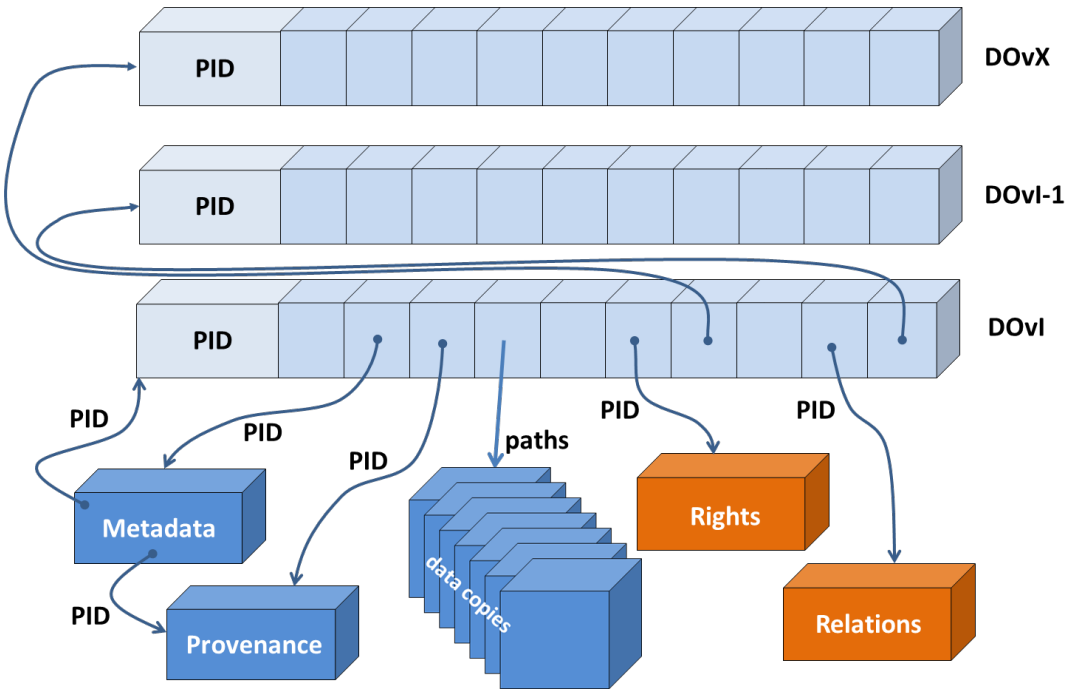


- CORE COMPOSITION 2 OR 3 INHERITED INTO NEW COMPOSITION
- CORE COMPOSITION COULD RUN ON SHARED INFRASTRUCTURE
- CORE COMPOSITION COULD BE RUN IN PRODUCTION MODE (LOCKED DOWN) OR OPEN MODE (EXPERIMENTATION WHERE EXPERIMENTAL COMPOSITION X CAN DROP IN EXPERIM TYPE SCHEMA

- We have seen that RDA has n=2 shared services around PIDs, can it go to n=3

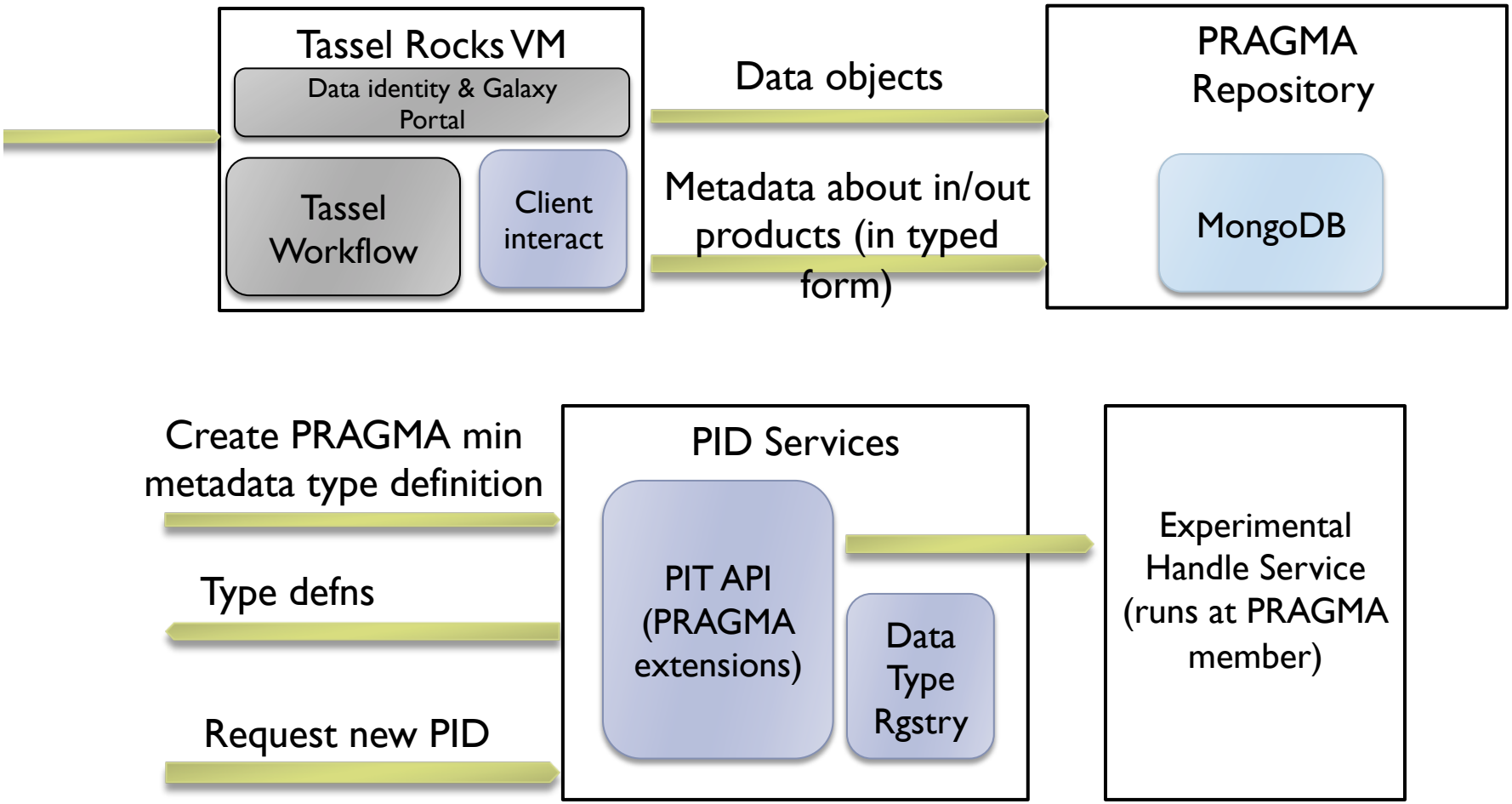
- Can it get to agreement around minimal PID attributes?

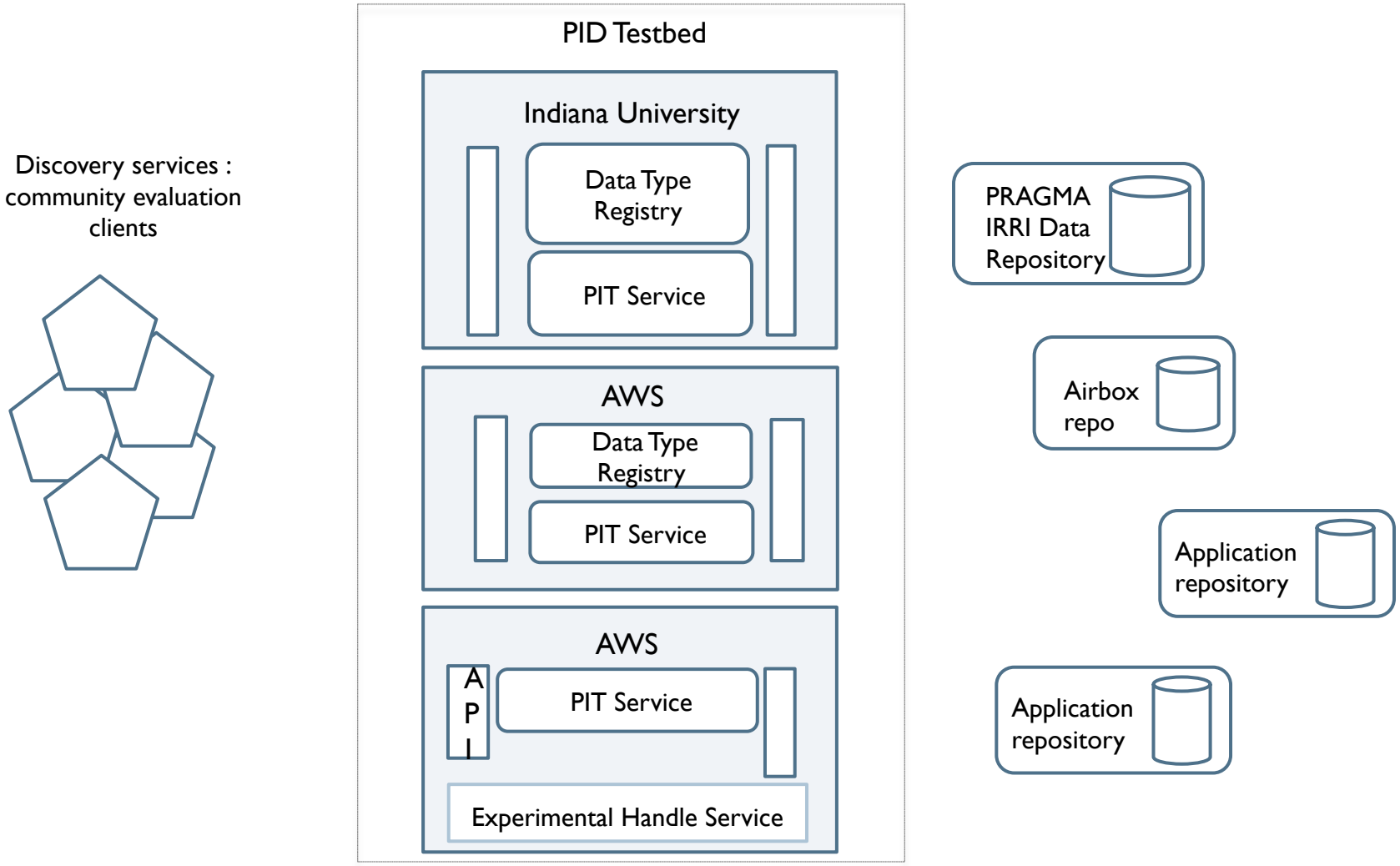
Example PID minimal metadata



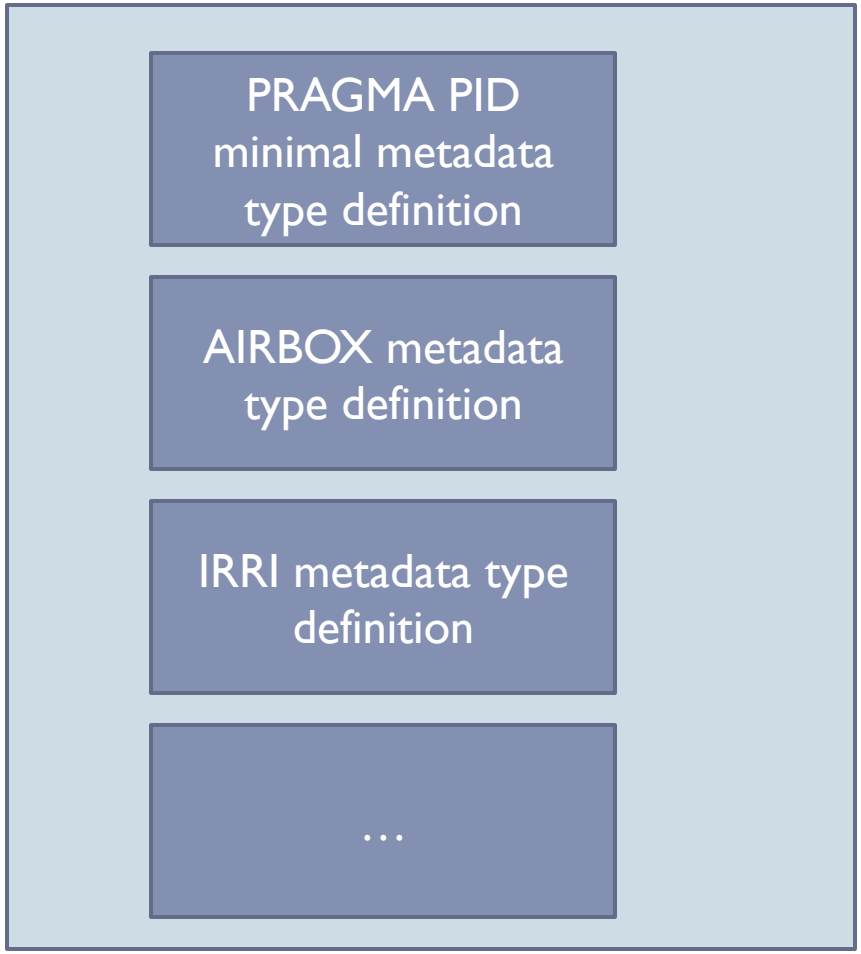
- Define PRAGMA defined definition (for experimental purposes … works for rice genomics, Airbox, weather data repository …

PRAGMA Rice Genomics: major components




Discovery services : community evaluation clients
PID Testbed
Indiana University
Data Type Registry
PIT Service
AWS
Data Type Registry
PIT Service
AWS
API
PIT Service
Experimental Handle Service
PRAGMA IRRI Data Repository
Airbox repo
Application repository
Application repository

Data Type Registry holds type definitions (aka, schemas for both the PID minimal metadata defn of a community plus metadata)



Deployment Diagram

PRAGMA Data Identity Service

- Data Identity Server
- PIT Service
- DTR Service

PRAGMA Data Repository

- Repository Rest API
- MongoDB DataBase

163.220.55.22
AIST Data Identity Cluster, Tsukuba

Data Identity GUI

Galaxy Workflow Server

- Tassel 5 Workflow
- Data Identity Tassel Client

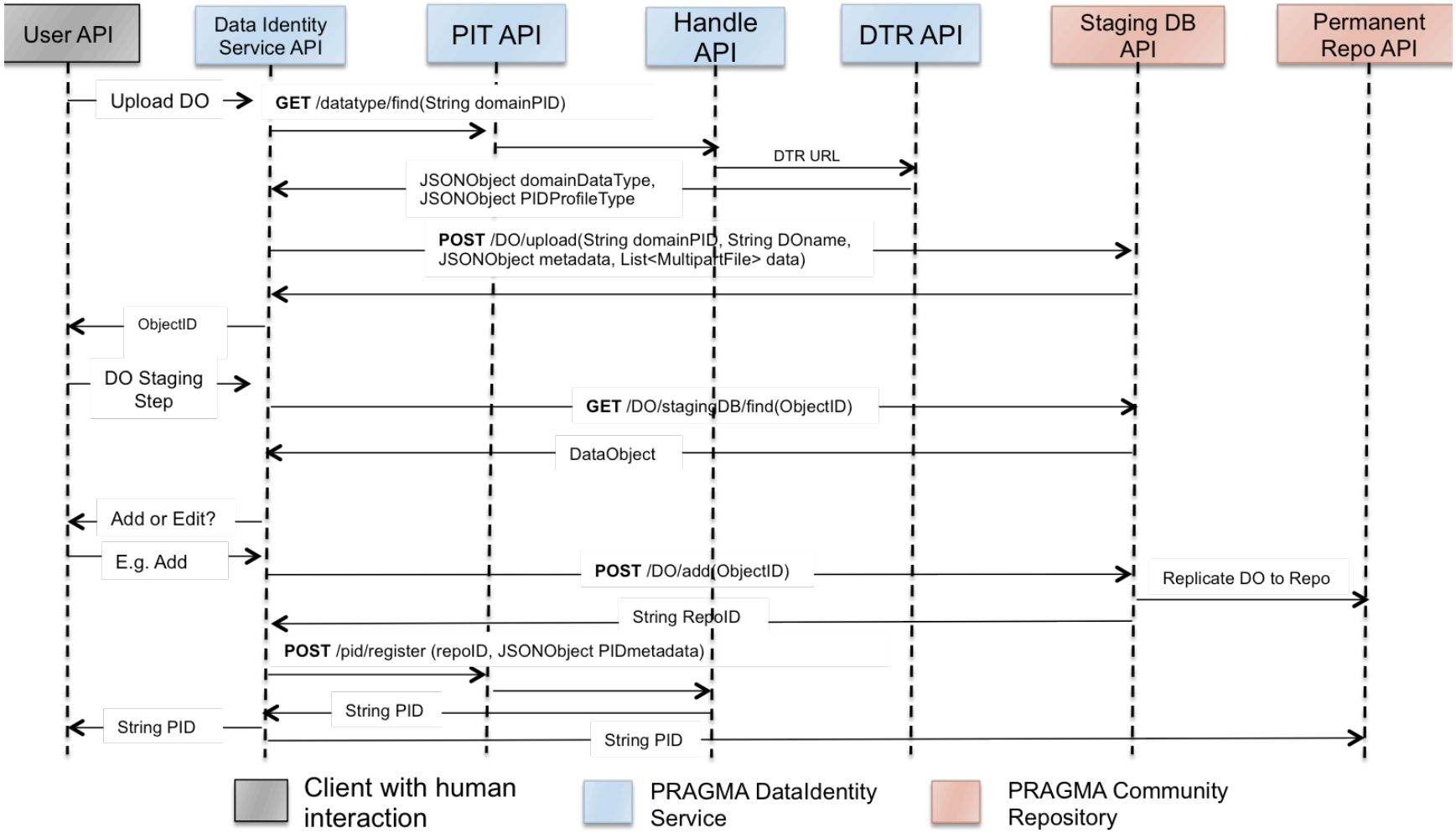
198.202.88.53
Rice Genomics Workflow VM
SDSC

Handle.net Service

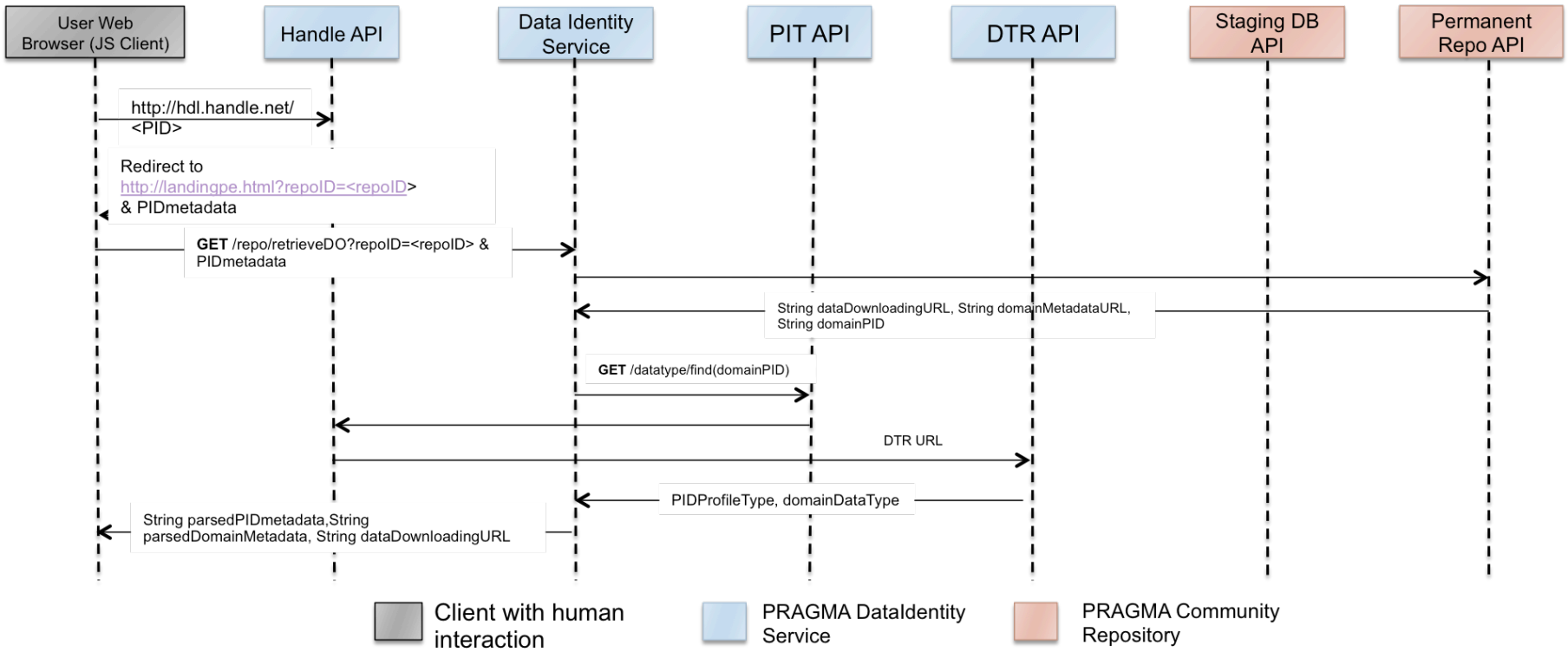
- Experimental handle V8 instance

38.100.130.12
PID Service Instance
CNRI, Virginia, US

DO Upload Timeline Diagram



DO Retrieval Timeline Diagram



Middleware Service Timeline Diagram