Comparison PID Kernel Info Provenance to Research Objects

Yu Luo, Isuru Suriarachchi, Beth Plale
Indiana University Bloomington, USA

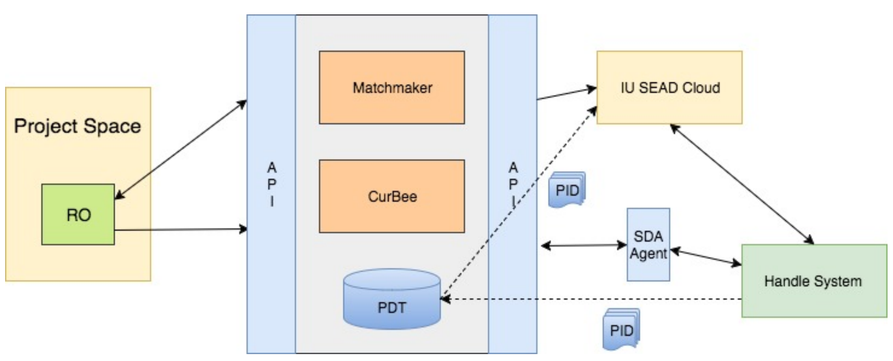
Motivation

- SEAD publishing pipeline is a cloud-based service for publishing digital data objects to one of a selected set of repositories.
- Publishing pipeline is built around the Research Object (RO) as the publishable entity.
- SEAD publishing services assign one DOI to every RO.
- But RO's can frequently be made up of any number of heterogeneous smaller objects.
- Data objects inside an RO have IDs that are not necessarily globally unique.
- RO is described by an OREmap file which
- We compare the existing solution with a solution that 1.) Replaces the DOI and local ID assignment scheme with Handles. 2. Replaces mapping of ORE with minimal set of metadata (called PID Kernel Information provenance) [3].

SEAD Publishing Services

SEAD publishing services:

- ***Project space:*** front end where user curates data objects and gathers select data objects into a curated publishable unit
- ***Curbee:*** ingest pipeline – carries out validation before publishing to a repository.
- ***Matchmaker***: selects a repository to which to publish
- ***PDT:*** a repository of profiles about (p)eople, (d)ata, and (t)hings (aka, repositories)
- ***IU SEAD Cloud:*** lightweight repository for storing Research Objects

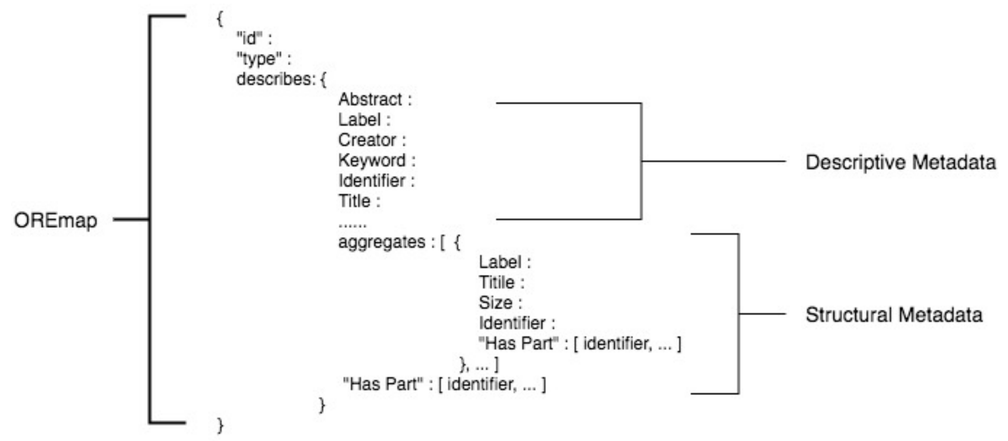


PID Kernel Information extensions to SEAD:

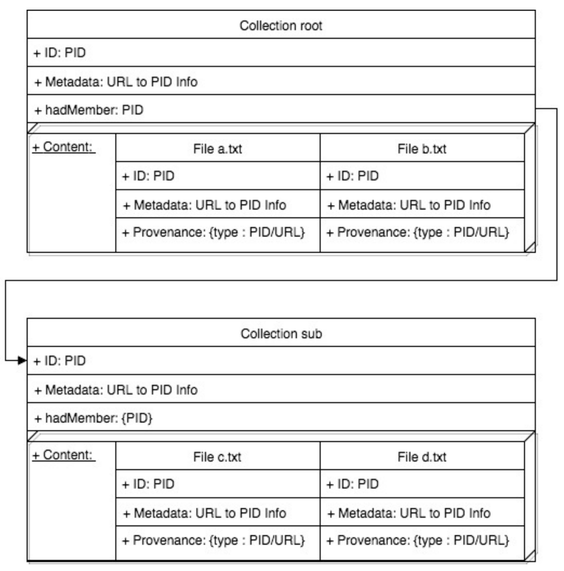
- IU SEAD Cloud invokes ***Handle System*** prior to deposit to create new handles
- IU SEAD Cloud records ***W3C Provenance*** into PID Kernel Information drawing on the provenance graph in the OREmap
- ***Provenance in PID Kernel Information*** serves as backbone graph connecting ROs.

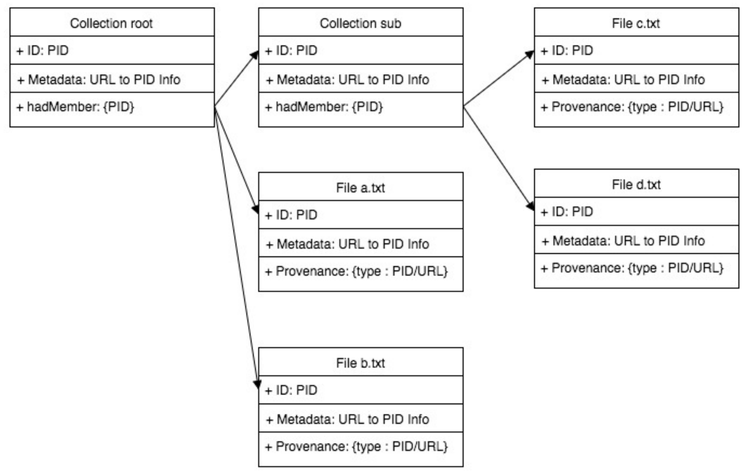
Comparative Study of Three Options A-C

A. OREmap file for an RO contains metadata including provenance and non-provenance information. Embedded metadata stores semantic and structural metadata alongside the data.



B. Embedded Model. Collection has direct reference to its contents. That is, relationship between file and collection embedded as part of collection metadata.





C. Reference Model. Collection has pointer to its members.

Study Metrics

Our study compares provenance-driven PID Kernel Information approach to SEAD's existing ORE approach. The following questions are addressed:

Is conversion from SEAD ORE representation to provenance PID Kernel Information representation lossless? That is, can provenance-driven PID KI completely represent a RO? Can RO be reconstructed from just possessing top level PID?

What are performance overheads of each of three approaches A-C?

What are strengths and weaknesses of each approach?

Funded in part through the National Science Foundation under grants 1234983, 1659310, 0940824, and grants from Microsoft Research for Azure cycles and the McArthur Foundation through Research Data Alliance/ US.

Broader Impacts

- Release PID option as an enhancement to IU SEAD Cloud
- Use enhanced IU SEAD Cloud to publish data to Azure as part of SEADTrain training effort
- Build out PID Kernel Information services as part of recently NSF funded, Research Data Alliance based, Robust PID testbed (RPID)
- Work with CENTRA partners to refine PID Kernel Information profile
- Work with CENTRA partners interested in evaluating the PID services

References

[1] B. Plale, *et al.*, SEAD: Lightweight Data Services for Sustainability Research, *JASM*, May 2014
[2] Robert Sanderson *et al.* Evaluation of OAI-ORE via large-scale information topology visualization. *9th ACM/IEEE-CS Joint Conf on Digital libraries*
[3] B. Plale, Power of PID Kernel Information, RDA WG meeting, NIST, Maryland, Dec 2016