Data Provenance as Killer App for PID Kernel Information

Presenter: Quan (Gabriel) Zhou

Quan (Gabriel) Zhou, Isuru Suriarachchi, Yu Luo, Beth Plale

Indiana University Bloomington, USA

4/13/17

Motivation

- Persistent Identifiers (PIDs) are globally unique IDs with a strong governance structure around them to ensure resolving authority. Our work uses the Handle System for PIDs. DOIs also use the Handle System but DOIs are better suited to publications instead of data.

- The Handle allows for a small amount of extensible defining metadata (that we call PID kernel information). This information, defined wisely, can enable an entirely new ecosystem of data services operating at Internet speeds.

- A killer app for PID Kernel Information is data provenance. By embedding only necessary and sufficient provenance into the PID Kernel Information, we take a big step towards universal provenance for data objects.

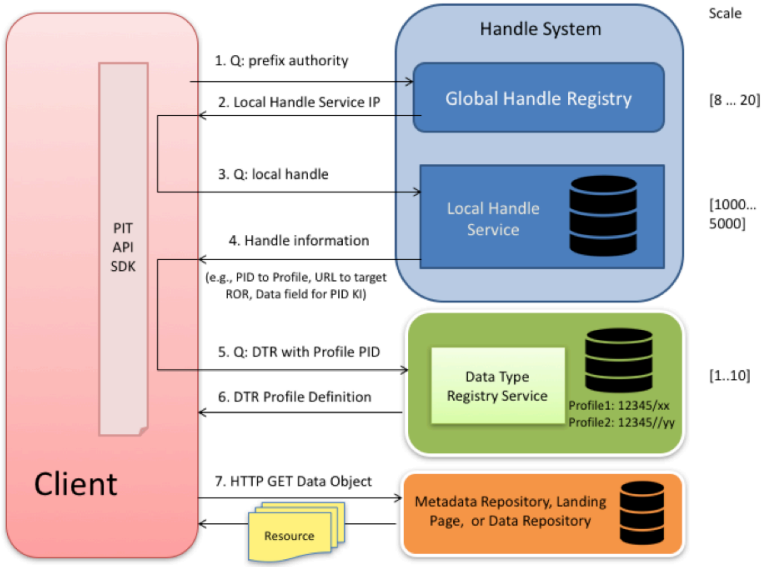
Use Case

"An internet scale provenance aware PID client receives a set of 100 million PIDs. It must first identify those PIDs which represents research data objects. It must make a judgment about whether a research object can be trusted or not."

Extended PID Architecture

Our RPID testbed, in progress, will

- **Store PID Kernel Information in Local Handle Service**
- Utilize **Data Type Registry (DTR)**, endorsed by RDA, to store the profile of PID Kernel Information
- Utilize **PIT API,** endorsed by RDA, as an SDK of tools for clients to interact with PID infrastructure
- Make the RPID testbed available for testing by communities



Provenance fields as part of PID Kernel information

	Type of Content	Content Format	Mandatory?	Explanation
1	wasDerivedFrom	IDENTIFIER	False	Transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
2	specializationOf	IDENTIFIER	False	Entity is of another shares all aspects of the latter, and additionally presents more specific aspects of the same thing as the latter.
3	revisionOf	IDENTIFIER	False	A derivation for which the resulting entity is a revised version of some original.
4	primarySourceOf	IDENTIFIER	False	Used for a topic refers to something produced by some agent with direct experience and knowledge about the topic, at the time of the topic's study, without benefit from hindsight.
5	quotationOf	IDENTIFIER	False	Used for the repeat of (some or all of) an entity, such as text or image, by someone who may or may not be its original author.
6	alternateOf	IDENTIFIER	False	Entities present aspects of the same thing. These aspects may be the same or different, and the alternate entities may or may not overlap in time.
7	hadMember	IDENTIFIER	False	A membership relation is defined for stating the members of a Collection.
8	externalW3CPROVDoc	URL	False	A URL referring to a W3C PROV document from an external repository.

Join us!

- Learn about how PIDs benefit your infrastructure.
- Contribute your use cases to developing the PID Kernel Information
- Access the RPID testbed to evaluate the RDA tools and our extensions
- Learn about RDA PID Kernel Information WG activity and contribute
- Contribute tools, services and clients to the growing tooling around PIDs in US and EU

- Contact: Beth Plale, Gabriel Zhou, Yu Luo

The End

Thanks!