Towards FAIR Open Science with PID Kernel Information

Yu Luo
School of Informatics, Computing and Engineering
Data To Insight Center
Indiana University

Strawman: Minimal metadata fields as part of PID Kernel Information

	Type of Content	Content format	Mandatory?	Explanation
1	PID	Handle	YES	Global identifier for the object; external to the PID Kernel Information
2	RDAKIProfileType	Handle	YES	Handle to the Kernel Information type profile; serves as pointer to profile in DTR. Address of DTR federation expected to be global (common) knowledge.
3	digitalObjectType	Handle	YES	Handle points to type defn in DTR. The type of the object (this should always be the same for this type of data, but would distinguish it from other data types). Distinguishing metadata from data objects is a client decision within a particular usage context, which may to some extent rely on the digitalObjectType value provided.
4	digitalObjectLocation	URL	YES	Pointer to the content object location (pointer to DO)
5	etag	Hex String	YES	Checksum of object contents
6	lastModified	ISO Date	YES	Last time of digital object modification
7	creationDate	ISO Date	YES	Date of digital object
8	version	String	YES	If tracked, a numerical version for the object

Strawman: 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.

Open science

Risk in defining open science too broadly

Open science must respect boundaries set by law or decency: licenses, copyright, human subjects privacy

Open Science increasingly connected to FAIR principles:

Findable

Accessible

Interoperable

Reusable

FAIR Guiding Principles

1. To be **Findable** any Data Object should be uniquely and persistently identifiable

1.1. Same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata

1.2. Data Object should minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects

1.3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent**

FAIR Guiding Principles

2. Data is **Accessible** in that it can be always obtained by machines and humans

2.1 Upon appropriate authorization

2.2 Through a well-defined protocol

2.3 Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object

FAIR Guiding Principles, cont.

3. Data Objects can be **Interoperable** only if:

3.1. (Meta) data is machine-actionable

3.2. (Meta) data formats utilize shared vocabularies and/or ontologies

3.3 (Meta) data within Data Object should thus be both syntactically parseable and semantically machine-accessible

FAIR Guiding Principles, cont.

4. For Data Objects to be **Re-usable** additional criteria are:

4.1 Data Objects should be compliant with principles 1-3

4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources

4.3 Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation

The Digital Object Architecture serves as base infrastructure only. DOA is silent on issues of modeling data objects themselves: their *content*, their *relationship to their own metadata*, and *relationship between data objects*

For object modeling we turn to FAIR principles and PID Kernel Information

Data Object Model based on FAIR principles

Data modeling questions address issues:

1) What goes into a data object?

2) Should a data object include its metadata or should the metadata be a new object or both?

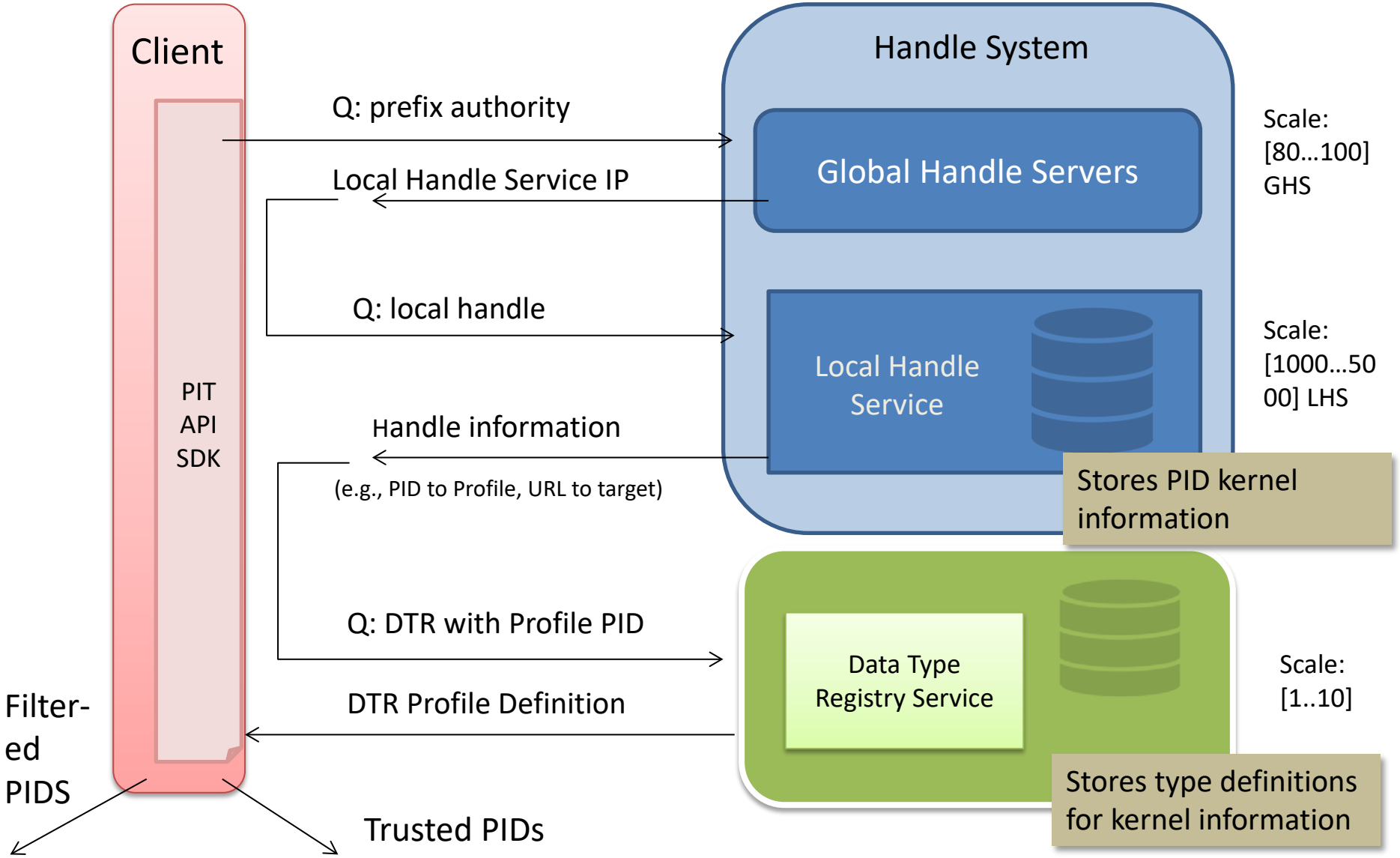
3) What kind of metadata should be considered?

4) What is the granularity of a data object?

5) Where does kernel information come in?

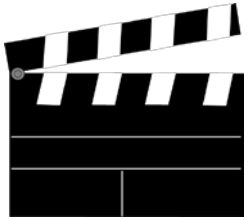
- Handle system allows key-value information stored to a Local Handle Server
  -- names a Data Object with name that is globally unique
  -- Data Object can be metadata, data or a digital proxy to physical object
  -- Is persistent over time

Handle resolution in a Digital Object Architecture



Client working with PID Kernel Information looks at each PID in list, accepts those that have:

-- Kernel Information profile stored in Data Type Registry (DTR),
-- That profile is associated with RDA (in some unspecified manner)
-- PID Kernel Information holds tiny amount of data provenance from which basic sense of trust is derived



What should go into the PID Kernel Information?

PID Kernel Information is a small amount of information stored at resolver (Local Handle Server) in PID record of a PID

Inspiration: take FAIR principles as guide: how far can PID Kernel Information aid in implementing FAIR?

Kernel Information is Cached

- By FAIR principle 1.1, a Local Handle Server is not a metadata repository so cannot serve as the authoritative source for any form of metadata for a data object

- Thus Kernel Information is cached copy of metadata that is stored and stewarded elsewhere

- *FAIR principle 1.1: Same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata*

Kernel Information for FAIR Accessibility

- By FAIR principle 2, Kernel Information conveys accessibility information thus making it easier for navigating direct data object access
- Includes privacy or legal restrictions on a data object that may limit access to, say the object's metadata alone.

*FAIR Principle 2. Data is **Accessible** in that it can be always obtained by machines and humans*

2.1 Upon appropriate authorization

2.2 Through a well-defined protocol

2.3 Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object

PID Kernel Information Summary

- Exploration driven by identifying and evaluating minimal information that can go into Kernel Information that can help make Data Objects FAIR and less dependent on the repository system to enforce FAIRness?
- Long term goal: Smart data objects
- Kernel information has potential to spawn new ecosystem of data services for smart data objects

RPID testbed

- Suite of software services for use by community
  - Data type registry (RDA)
  - PIT API (RDA)
  - Handle service
  - RDA Collection API
- Exploratory services
  - PID Kernel Information
  - Mapping CTS URNs to handles
  - Packaging for use by others
- Help and advice
- User advisory group

- Follow work at:
  - https://github.com/rpidproject
  - RDA PID Kernel Information Working Group
  - Reach us at rpid-l@iu.edu

Acknowledgements: this work funded in part by the National Science Foundation under grants 1659310 and 1349002