

Persistent Identifiers as Boundary Objects: A socio-geographical view of standards development and implementation

Mark Parsons¹ and Peter Fox¹

¹Affiliation not available

November 26, 2022

Abstract

In this poster, we begin to explore how socio-geographical considerations can inform the development of data infrastructure, notably Persistent Identifiers. PIDs have become largely accepted within the Research Data Alliance, W3C, and elsewhere as core elements of data infrastructure. Science is comprised of many divergent formal and informal viewpoints at many different levels with a need for generalizable findings. PIDs act as “Boundary Objects” (Star & Griesemer, 1989) — objects that are part of multiple social worlds and facilitate communication between them. They allow meaning to be understood in different contexts and are “plastic enough to adapt to local needs, . . . yet robust enough to maintain a common identity across sites. They are weakly structured in common use and become strongly structured in individual site use.” Boundary objects work to reduce local uncertainty without damaging cooperation. It is a question of re-representations across intersecting worlds not consensus. PIDs work to allow machines and humans to understand which digital object is in question (identity), what it is (type), and where it is (location). Each of these questions is surprisingly fraught and complex.

Persistent Identifiers as Boundary Objects

A socio-geographical view of standards development and implementation

Mark A. Parsons (@chutneyboy, parson3@rpi.edu) and Peter A. Fox (@taswegian, pfox@rpi.edu)

<http://tw.rpi.edu>



DEEP CARBON OBSERVATORY DATA SCIENCE

RDA RESEARCH DATA ALLIANCE

ALFRED P. SLOAN FOUNDATION

Bridging Heterogeneity

In this poster, we begin to explore how socio-geographical considerations can inform the development of data infrastructure, notably Persistent Identifiers. PIDs have become largely accepted within the Research Data Alliance, W3C, and elsewhere as core elements of data infrastructure.

Science is comprised of many divergent formal and informal viewpoints at many different levels with a need for generalizable findings.

PIDs act as *“Boundary Objects”* (Star & Griesemer, 1989) — objects that are part of multiple social worlds and facilitate communication between them. They allow meaning to be understood in different contexts and are “plastic enough to adapt to local needs, ... yet robust enough to maintain a common identity across sites. They are weakly structured in common use and become strongly structured in individual site use.”

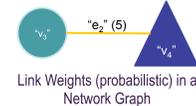
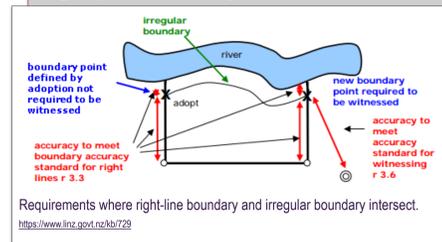
Boundary objects work to reduce local uncertainty without damaging cooperation. It is a question of re-representations across intersecting worlds not consensus.

PIDs work to allow machines and humans to understand which digital object is in question (identity), what it is (type), and where it is (location). Each of these questions is surprisingly fraught and complex.

Types and Boundaries

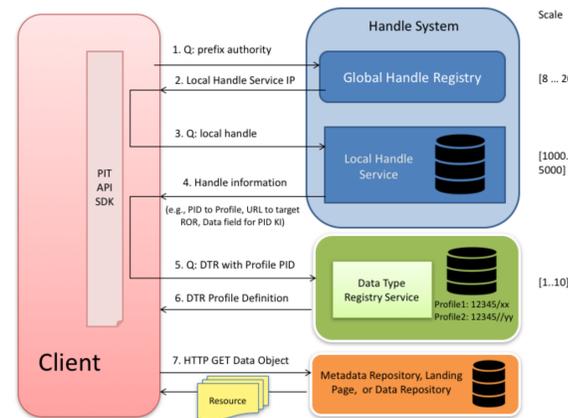
Types creates boundaries, but boundaries transcend types by highlighting both delineation and overlap.

- Category types are necessary but no one classification organizes the world for everyone.
- Standards remove diversity yet standards are intensely local.
- Geoscientists work to understand boundaries and groupings. Similarly boundary objects mediate between different groups. They do not provide a consensus view or common understanding of the different groups. Instead, they illustrate boundaries in both a geographic and mathematical sense that mark differences and areas of intersection.



Examples of intersectional boundaries

Schematic of PIDs and Types



Identity, Location, Name (and Persistence)

These are related but different concepts that are often confused when speaking of digital ‘identifiers’. IETF, W3C, and DONA try to sort out the confusion:

- **Uniform Resource Identifier (URIs)** (IETF RFC 3986): ‘identify’ a resource. A URI can be a locator, a name, or both.
- **Uniform Resource Locator (URLs)** (RFC17380): a subset of URIs that identify a resource and describe its primary access mechanism (its network “location”). IETF has deprecated “URL” and recommends URI with the declared access protocol.
- **Uniform Resource Names (URNs)**: location-independent, persistent identifiers for information resources and a type of URI. There are multiple URN namespace authorities, including ISBN, but currently no global resolution services. This remains an active area, and two proposed RFCs have been published in 2017 — 8141 and 8254.
- **Registered PIDs**, like Handles, DOIs, and ARKs, address the concepts of name and location separately albeit under one identifier string. The key is that these systems include a managed resolution service. PIDs can be URIs but the structure for maintaining the naming and location resolution service is what makes the identifier “persistent” as both a name and a locator. Handles are described in RFCs 3650, 3651, and 3652.



Considerations Going Forward

- Don’t underestimate the complexity of “identity” (or class or type). Make it a principle to consider multiple, non-privileged perspectives.
- It is not a question of a least common denominator but rather a fundamental, simultaneous, n-way translation requiring a holistic view to understand. Pose many questions of thirdness to get a ‘good-enough’ intersectional model.
- An inclusive culture can only help — props to AGU’s new Integrity and Ethics Policy (<https://ethics.agu.org>).
- “Draft” standards are more complete than the name implies, yet standards are always a work in progress and that work needs to be integrated into routine research activity.
- Adoption and enhancement of boundary objects occurs through durable, arful collaboration and situated learning
- Adoption changes both the standard and the adopting organisation and can ripple to other organisations.

What is “Band Practice”?

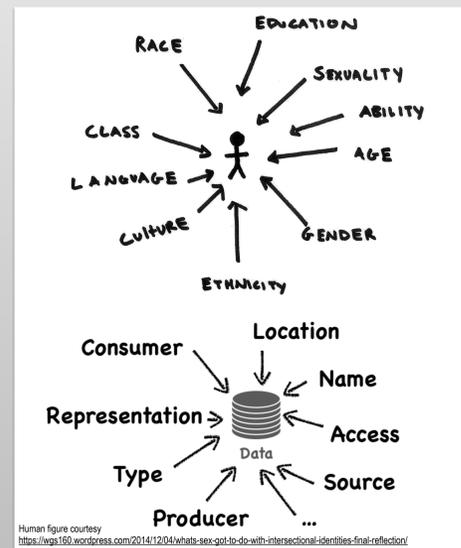


Intersectionality

“Not just for social justice warriors!” – A. Levy

- Intersectionality requires that we examine multiple variables at once because they often work in association with each other.
- It is particularly concerned with formations of social identities and how “challenges faced within interlocking systems of oppression of race, class, gender, and sexuality could not be solved by mono-categorical solutions.” (Collins, 2014).
- The same approach applies to other forms of identity and requires reconciling meaning across all types of difference.
- Not a moral question but a balance of trade-offs.

Examples of Human and Data Identity



Pragmatics, Technology, and Adoption

PID Kernel — minimal information stored in the PID record. The goal is to curtail rapid growth of types and realize a long-term core of “universally” understood information.

- PID Kernel needs to be a problem of Peirce’s firstness as interpreted by Gahegan & Adams (2014) (defining firstness has risks) and reference minimal secondness, i.e. strict syntax and minimal well-controlled semantics. Thirdness (context per Latour) must be accommodated but not stored in the kernel.
- Model from both the data consumer perspective and the producer perspective to better describe the how, who, and why in addition to the what, where, and when.
- Avoiding path dependence and “ASCII-like Imperialism” (Pargman & Palme, 2009)



Proposed Principles for a PID Kernel

PID Kernel Information:

1. enables machine-actionable services that operate at Internet scales. (minimal)
 2. is stored locally not referenced (efficient, small, fast).
 3. in a local store is not authoritative but a duplicate of authoritative version elsewhere (controlled).
 4. attributes are those that have slow rate of change. (no human interaction on updates)
 5. content record are property of data object owner or delegate (repository). Because the PID KI is redundant by (3), updates are allowable only by owner or delegate.
- (courtesy RDA PID KI WG: <https://www.rd-alliance.org/groups/pid-kernel-information-wg>)

PID & Type Adoption and Use Cases

- PID kernel development will continue, but we must start somewhere.
- “The Deep Carbon Observatory (DCO) is a global community of multi-disciplinary scientists unlocking the inner secrets of Earth through investigations into life, energy, and the fundamentally unique chemistry of carbon.”
- DCO organizes a Data Portal — a managed digital object identification, object registration, and metadata management service providing discovery and access to diverse data for the DCO community.
- The DCO-ID (a Handle) is at the core of the portal functionality and is used to identify and locate publications, people, organizations, instruments, datasets, sample collections, keywords, conferences, etc. across many institutions.
- DCO has adopted and is exploring the adoption of several RDA Recommendations based on PIDs (see Poster: IN43A-0067). We are also active in the PID Kernel WG.
- Because the DCO-ID is so multidimensional and because we are exploring use cases well beyond scholarly communication, this should help address issues of identity and intersectionality.
- Past adoption suggests that it may occur differently than the developers planned (e.g., a triple store instead of RDB).