Thoughts on PIDs for Facilities and Instruments
by Mark Parsons

Bear with me. I'm going to start out at a rather meta level coming from a social science perspective and then I'll try and bring that down to something more concrete. This is what I do. I try to understand the underlying social and theoretical issues we are trying to solve in order to guide more specific action.

So first a basic assertion I think we will all agree with: A network of FAIR data and services and associated credit *requires* that research objects are unambiguously identified and located. We do this with persistent "actionable"* identifiers —PIDs: Unchanging names of entities (URNs) with a mechanism of resolving this to a location or access point. In other words, it's a registry. Nate's metaphor of a (warehouse) inventory system is dead on.

To keep track of assets, these the two functions—name and location—must be addressed and maintained, and location in particular has a a fundamental, sustained, institutional component independent of technology.

And this is how we've always done it. Registries are arguably a fundamental basis of civilization. More than 5000 years ago, people started recording assets (grain stores, property, etc.) and this formed the basis of writing and myriad social structures.

This leads to my 2nd assertion: registering things is an act of power. It is those who have power that get to decide what is important, what gets a name, and what gets registered and tracked.

Benedict Anderson made this argument really well back in 1982-3 in one of the most cited books in social science called "Imagined Communities". He was discussing what defines a "nation" (a relatively modern concept). I especially like the chapter entitled "Census, Map, Museum":

I quote: "These three institutions … profoundly shaped the way in which the state imagined its dominion – the nature of the human beings it ruled, the geography of its domain, and the legitimacy of its ancestry." Note these three things (CMM) are essentially registries.

When we look over the history of PIDs, we have seen power competition between various registries and their proponents. Often friendly competition but not always. I would argue this is the heart of the ARK vs. handle (or DOI) debate, for example. I think this is what Shelley meant yesterday when she called for a come-to-jesus moment. We all need to atone, forgive, come together and agree on respective roles and responsibilities. And its not just the registries that need to do this.

So a 3rd assertion: Identity is fraught and it is best defined through intersectionality. In other words, a thing is defined by its relationships to other things. This can diffuse some of the power.

We typically think of intersectionality in terms of people. I can be identified with my passport or ORCID (or race or gender), but if you really want to understand who I am, you need to understand my relationships to other people, institutions, and cultures as well as the work I have done. I think intersectionality applies equally to other things including facilities and instruments. They are defined by what they do. To what. For whom.

I think this has been a big theme of this meeting. As Matthew just said, verbs are important. The edges in a graph may be more important than the nodes. As Anita said on the first day "PIDs are best when they work together". PIDs are linkers. We saw this in the talks yesterday -- from David Elbert and Caterina's specific examples, to the graphing efforts of Neil and David Hart, and especially Ted's talk about DataCite's 36 (underutilized) relationship types. I think the FAO and GBIF examples Ted discussed highlight approaches that do relationships well.

So this brings me to a central question which has run throughout all three workshops. When do we need a PID and when do we need metadata? I'm leaning more toward metadata.

As Nate said, no one PID will solve all problems. And even then, you need to use them well according to defined leading practice. As Neil pointed out we need to understand what are the questions we are trying to answer and are PIDs worth the lift for the use case?

My take is that PID metadata should be pretty skinny, but it should emphasize relationships. We know that PID metadata is often just the required fields. So let's require at least two relationships -- the resolution to the object (or landing page) and at least one more. I don't really care what that additional relationship is as long as it is defined. I am quite fond of hasMetadata, though. And that could be repeated. Multiple organizations can and do maintain metadata about a thing. Centralize name and location. Distribute metadata.

To use data as an example, back in the day, when the DataCite metadata schema was being defined, I never really understood why they wanted to go beyond the basics to include things like parameters or temporal coverage. I thought they should instead just link to the nicely structured metadata held by the repository. They should also link to other sources of documentation, like a paper describing the sampling protocol. These extra things can all be maintained independently to address different uses (distributing the power) while the PID focusses on its core functions -- name and location.

Finally, we have the issue of maintenance. As I noted at the beginning, PIDs are only as persistent as their maintainers. We need to be clearer in the Recommendations who has responsibility for maintaining what. Be it a PID or metadata or other things. As Maria from DataCite said: "metadata quality and completeness is a collective challenge and responsibility!" This gets into issues of governance as the data professionals group pointed out yesterday. It also means agreeing on some semantics especially about relationship types. Talk to Doug Fils. This is complex and I could say more but I've talked more than enough already.

So I'll summarize:

Registering assets is essential, and relevant information must be maintained. Registering things is an act of power. Let's diffuse or distribute that power. We can begin to do that by recognizing that what is a thing is defined by its relationships.

That leads me to propose some next steps:

1. Lead with the use cases and be as specific as you can. (the Recommendations do that, but it's clear we need more specifcs and examples)
2. Clarify as best as possible when PIDs address a problem vs when metadata addresses the problem. Focus on the verbs, the defined relationships across objects.
3. Clarify who is responsible for maintaining what and how that should be (loosely) governed.