

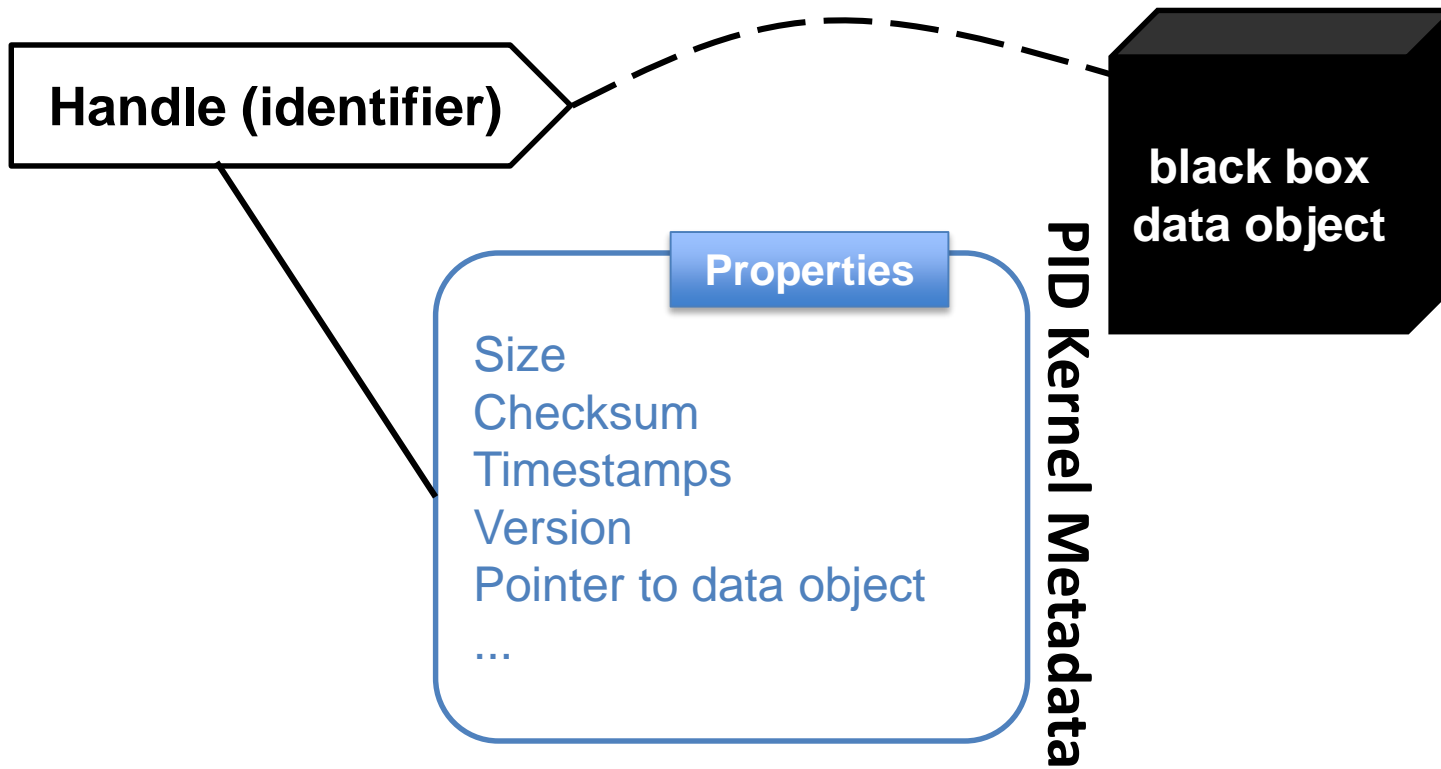
# Power of PID kernel information

Beth Plale

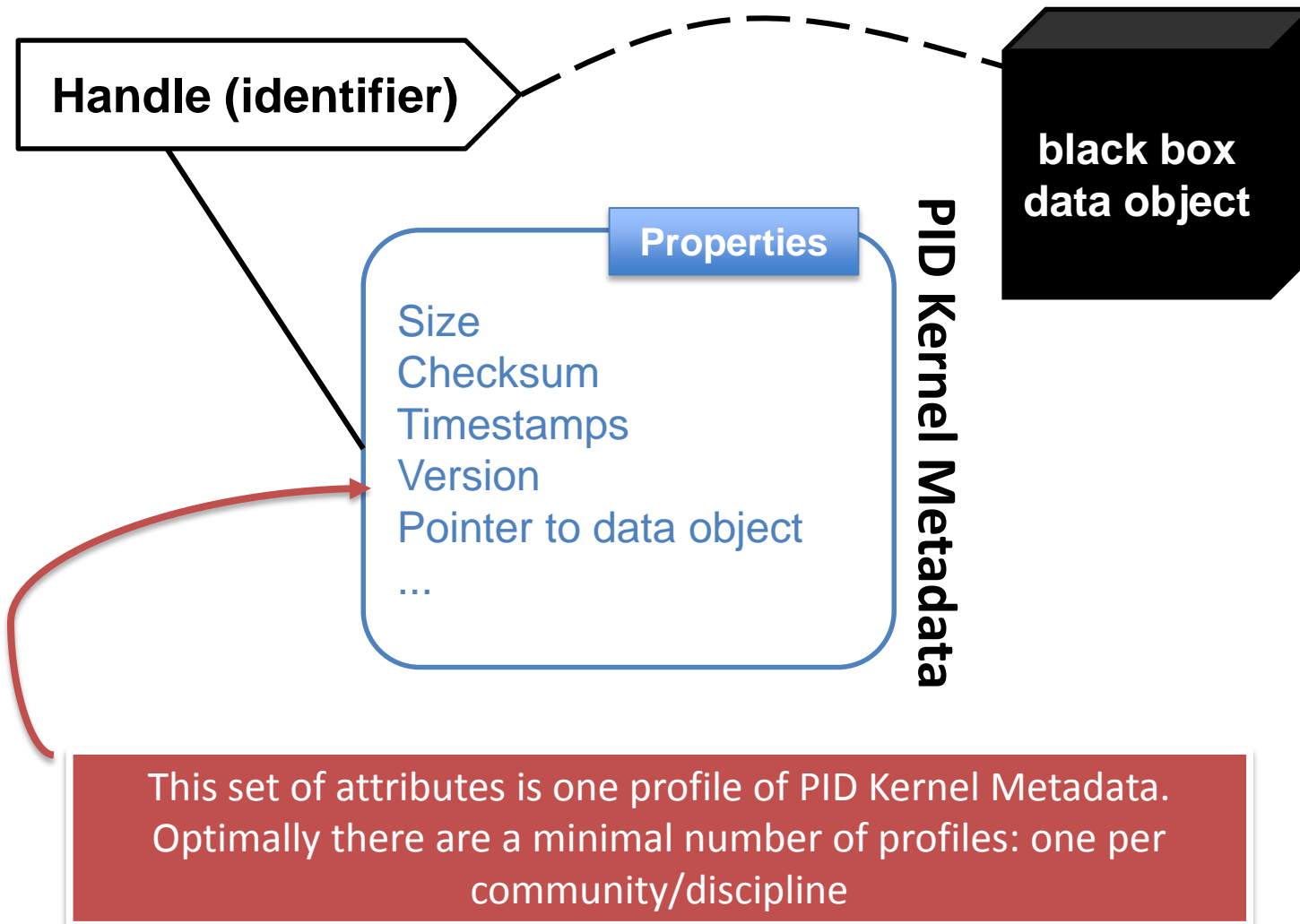
Indiana University

Bloomington, Indiana USA

# Conceptual model



# Conceptual model



# What a profile is

- Base assumption: There is minimal set of information associated with each PID
  - Names for this metadata: *kernel*, casing, gateway metadata
- Kernel metadata should be useful to
  - maintainer of Data Object,
  - discovery clients, and
  - Data reuse research
- Each user community may design their own profiles.
  - No single size fits all – but full benefits realized when set of profiles is minimal

Size  
Fixity Key  
Timestamps  
Data Provenance  
...

Already in place for evaluation is a PID Testbed:

- Runs on ***PRAGMA Testbed***
- Tools:
  - Drawn from Research Data Alliance (RDA) Recommendations
  - Persistent Identifier Types API and
  - RDA Data Type Registry



# How get engaged

- Identify small dataset of yours (e.g., Airbox) to play with
- We'll help you use PID assignment services to assign PIDs for the data in your dataset
- Work with us to define 1-2 kernel metadata profiles
- Work with us to grow this out

# Who

- Beth Plale, IU
- Jay Combinido, W ASTI
- Krisanadej and Mullica Jaroensuttasinee, Walailak
- Dinh, Van Dzung, VNU
- Jose Fortes, UF

# What

- Datasets
  - Hathitrust Library
  - iDigBio
  - Weather data (ASTI)
  - Ecological data (Thailand)
- PID service prototype
  - Generate PIDs
  - Integrate with metadata systems (e.g. IRODS)
  - Profiles
- Centra Webinar
- Google folder/doc, Facebook group



# When

- April 10-14
- 10-12 April 2017 -- CENTRA All-Hands meeting on Smart and Connected Communities
- 12 April 2017 – CENTRA/PRAGMA SUNTOWNS (Smart University TOWNS) Workshop
- 12 April 2017 -- GLEON Lake Modeling Workshop
- 13-15 April 2017-- PRAGMA 32 meeting on Internet of \*People and Things

# What

- Datasets
  - Hathitrust Library -- Beth
  - iDigBio -- Jose
  - Weather data (ASTI) -- Jay
  - Ecological data (Thailand) – Mullica and Krisanadej
- PID service prototype – Beth et. Al.
  - Generate PIDs
  - Integrate with metadata systems (e.g. IRODS)
  - Profiles
- Centra Webinar – Beth et al.
- Google folder/doc, Facebook group -- Gabriel

# Benefit to your project

- You can claim that you evaluated two RDA recommendations and were part of PRAGMA/CENTRA contributions to active RDA effort (third recommendation)
- Provenance in PID kernel metadata is a killer app; we have opportunity to be global leaders here.

I encourage you to reach out to me  
[plale@indiana.edu](mailto:plale@indiana.edu)  
*[www.linkedin.com/in/bethplale](http://www.linkedin.com/in/bethplale)*



- Persistent IDs (PIDs)

are globally unique

name a data object (metadata or digital data, not physical object)

are persistent



# For a PID to be Persistent:

- Relationship between ID and object that it references needs to persist
- PID system should be scalable, governed, and governed through an open body
  - Options for science:
    - Digital Object Identifier (DOI)
    - Handle system






Making research better by enabling people  
to find, share, use, and cite data




DataCite slides credit: Laura Rueda, DataCite. Presented at RDA PID Training Garching Germany Aug 2016

# Our mission

- DataCite develops and support methods to:

-  locate
-  identify
-  cite

data and other research objects to:

-  establish easier access
-  increase acceptance
-  foster reuse





DOIs find most use in  
publishing final digital  
products to public domain

...

But ... research world needs  
PIDs for much more than final  
digital products



# HathiTrust Research Center: Provisioning Analytical access to massive digital library

- 14,724,583 total volumes
- 7,354,226 book titles
- 404,301 serial titles
- 5,153,604,050 pages
- ~39% in public domain

Looking for repeatable page and phrase level access to content in book titles.

Manage 10 billion DOIs? ... ahem, no.

plale@indiana.edu

## Literary Geography at Scale

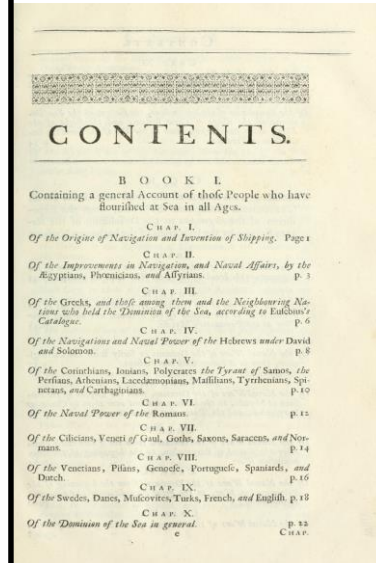
Matthew Wilkens, University of Notre Dame

With the help of natural language processing, Dr. Wilkens will extract and geocode place names in nearly eleven million volumes from the HathiTrust, including twentieth and twenty-first century texts. This project to geolocate world literature, supported by a 2014-2015 American Council for Learned Societies Digital Innovation Fellowship, is one of the largest humanities text-mining projects to date.



- Pilot analysis completed and named entity extraction run on 10,000 randomly sampled volumes.
- Pilot took 55 minutes of processing time on Indiana University's high-performance computer, Big Red II, using an 8-node cluster.
- Next step: Process the entire corpus!

Collaborating with scholars





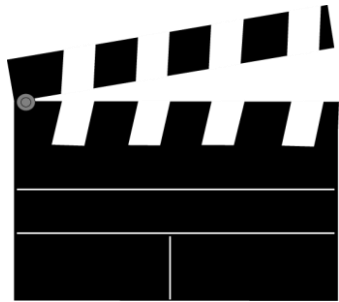
# Alternative?



Uniform adoption of PID  
use across all research  
stages and disciplines can  
stimulate an entire  
ecosystem of discovery  
services for research data



# Imagine a world where PIDs identify just about everything:



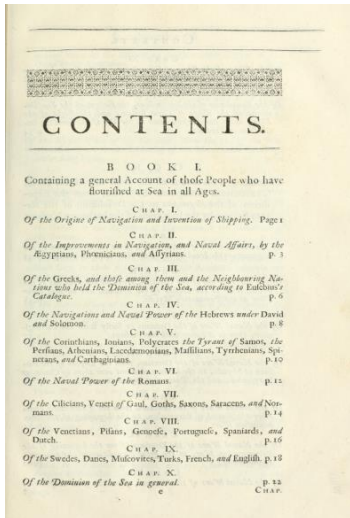
-> Internet of Things

-> Movie clips

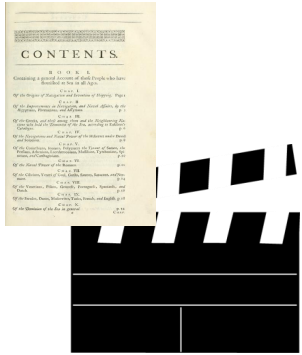
-> Smart city sensor data

-> Pages from digitized books

-> Baby food containers



Go on to imagine an Internet-scale data client that is handed a list of a billion IDs.



Case 1: How does the client quickly sift through the list to find the research data?



Case 2: When client winnows list down to research data, how does it then quickly discard the fakes?

Solution:

PIDs with kernel information  
that is *actionable*,  
*automatable*, at  
*Internet scale speeds*

