Persistent Identifiers

Daan Broeder

Max-Planck Institute for Psycholinguistics

DAM-LR

Identifying what?

- Things you want to identify in a unique and persistent way.
  - Web accessible resources
  - Services producing resources
  - Lexicon entries, concepts from concept registries,…
  - …

…… and want to access

Resource Identification on the Web

- URI: Uniform Resource Identifier

  generic name for all ways of identifying resources on the Internet

  - URL: Uniform Resource Locator
    - http://www.mpi.nl/IMDI/metadata/IMDI.xml
    - When used as an identifier, the URL is both name and location of the resource
  - URN: Uniform Resource Name
    - urn:ISBN:0262531283
    - Namespaces should be registered, isbn, ietf:rfc, mpeg:mpeg7:schema,…

DAM-LR

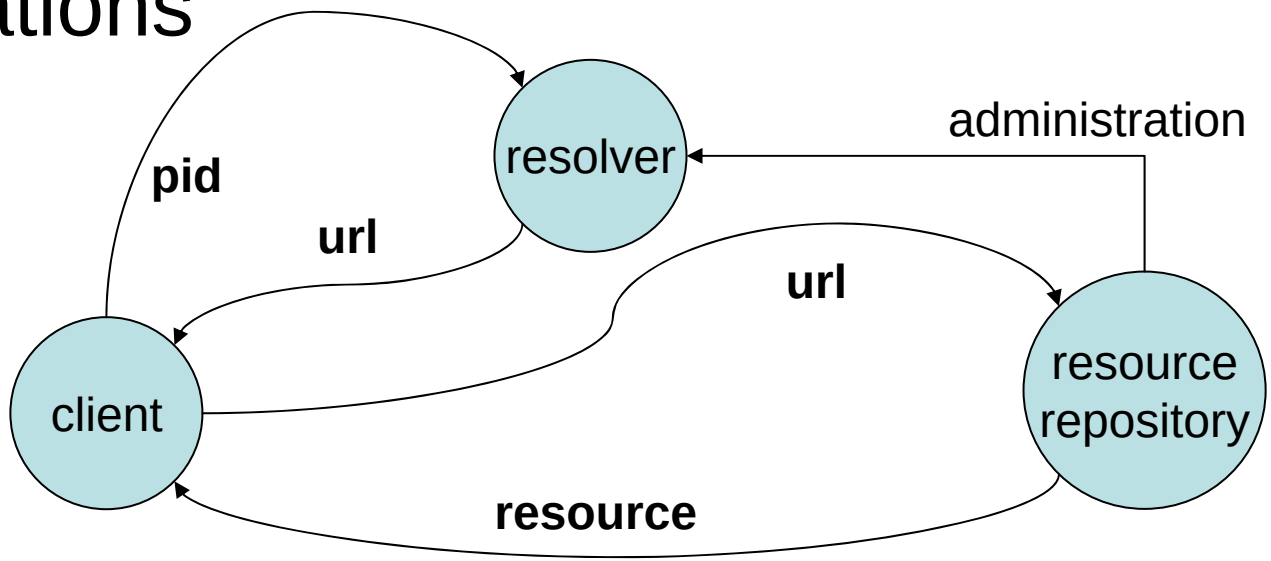
Using URLs as Identifiers

- Embedded URLs become actionable in many tools.
- However:
  - When resources are moved dead links result (link rot). Unless you succeed updating all referring documents.
  - The URL string may hold meaning e.g.

    http://www.mpi.nl/data/corpora/non-evaluated-resources/house.wav

Persistent Identifier Systems

- Separate resource name from resource location
- Resolver system to translate names into locations



Persistent Identifier Systems

- Associate extra information with the PID
  - URLs of copies
  - Metadata
  - Authentication system info
- High availability, scalable, …

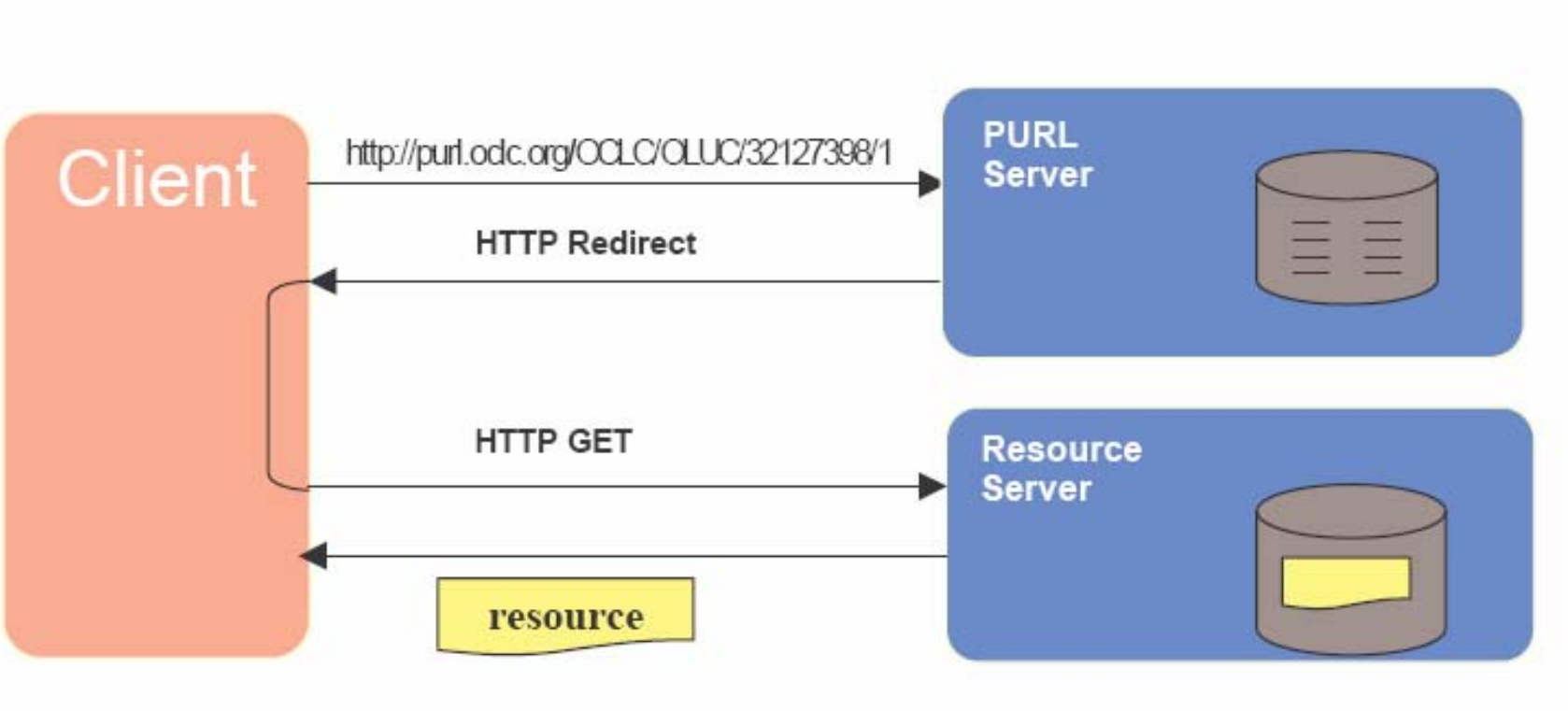
Existing PID Systems (not an exhaustive set)

- PURL: Persistent URL, based on http redirect
- HS: Handle System, full fledged pid system
  - DOI: HS implementation
- ARK: Archival Resource Key

Persistent URL (PURL)

- Developed by OCLC (online computer library centre, 1995)
- Format:
  - http://purl.oclc.org/emls/texts/libels
    | prot | resolver | asset name
- Works by HTTP redirect
- Purls are directly actionable
- Binds only one single URL with the identifier, no metadata
- Use of central purl service is free. Free server software to create your own service.
- 669733 purls registered, 378307173 resolved, but possibly repositories run their own resolver.

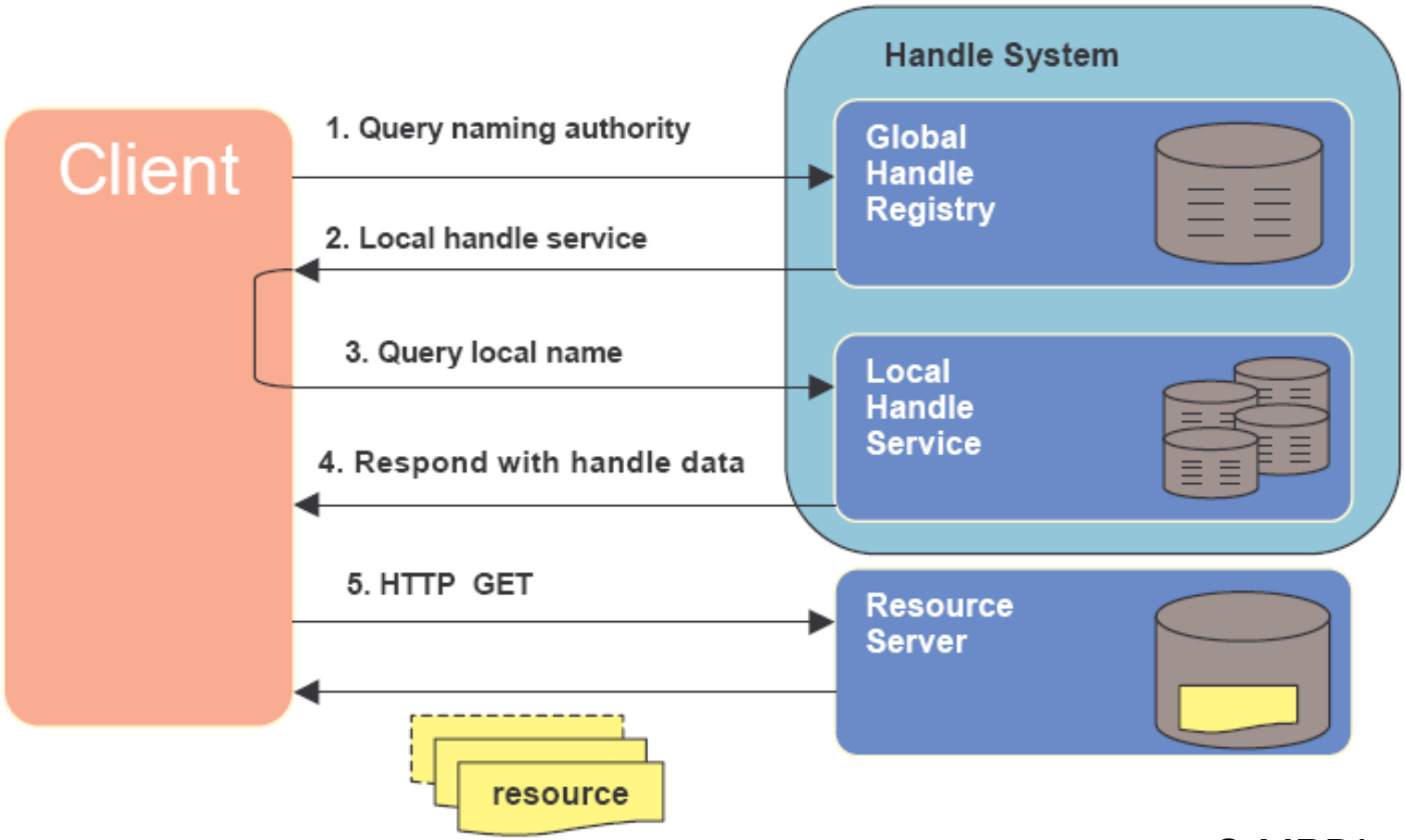
PURL Resolution



Handle System (HS)

- Developed by CNRI (Corp. for National Research Initiatives 1995)
- Format:
  - prefix/name, prefix obtained from GHR
  - 1839/00-0000-0000-0000-0000-4
- Distributed resolving system like DNS
- Can resolve identifier to any kind of associated info
  - (multiple) URL
  - metadata
- Scalable, secure, independent of the http protocol
- Actionable by using a plug-in or URLifying the handle
  - http://handle.net/**1839/00-0000-0000-0000-0000-4**
- Free software, small fee for registration with global HS
- Publishers are an important HS user community with DOI

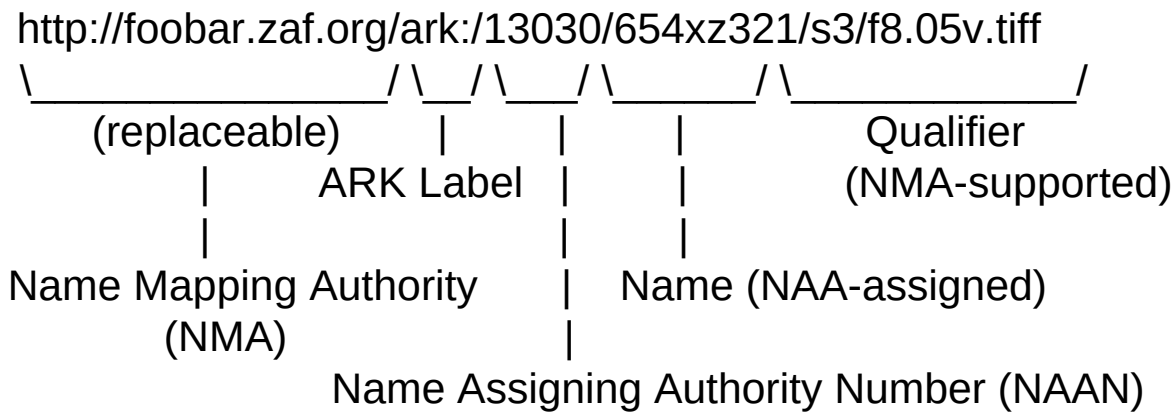
HS Resolution



ARK: Archival Resource Key

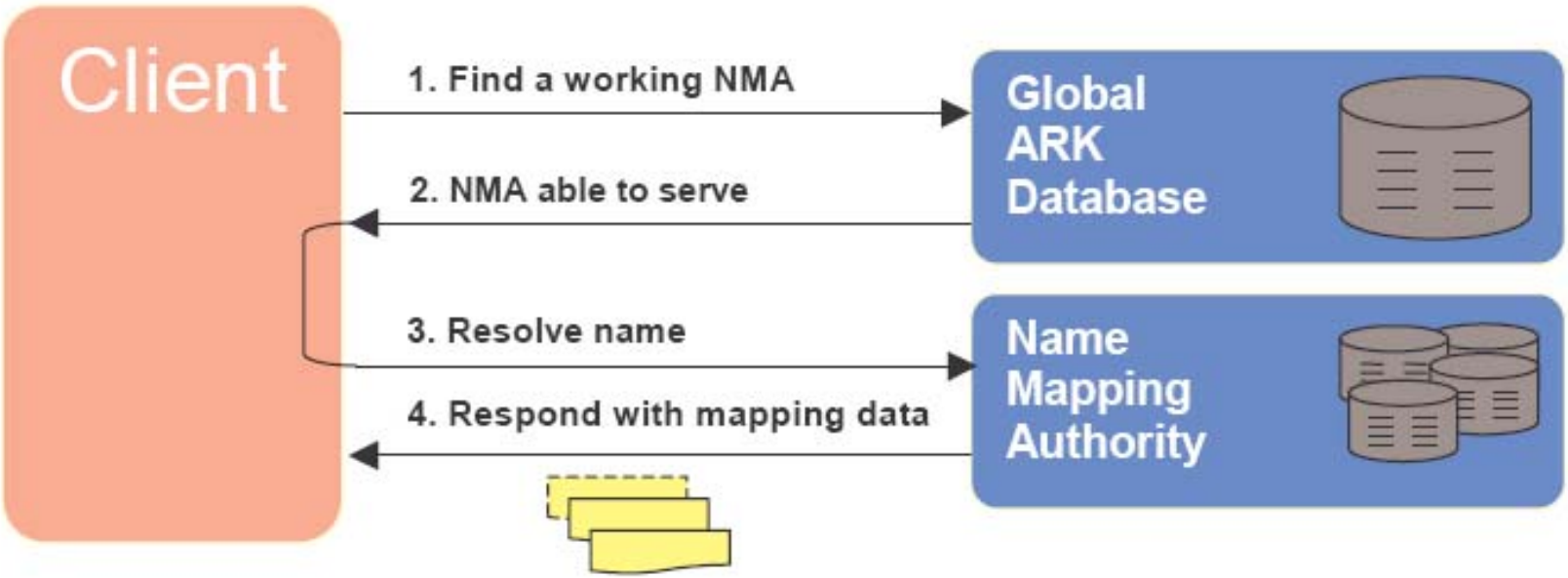
- Developed by CDL (California Digital Library, 2001)
- Format:
  - NMAH:ark:NAA:Name[Qualifier]
  - http://ark.cdlib.org/ark:/13030/ft4w10060w
  - Qualifier -> complex/hierarchical objects + variants
- Resolves/Delivers
  - Location
  - Metadata (ERC)
  - Commitment statement : *Duration of association, Duration of availability, Version policy*
- Directly actionable if NMA provided with the ark.
- Enforces policy
  - No semantics in identifier strings
  - *Type of metadata records*
- *Some describe it as work in progress*
- *Software: sketchy, contradictory info, maybe no standard NMA package* yet

ARK format



- 12025 National Library of Medicine
- 13030 California Digital Library
- 13960 Internet Archive
- 27927 Portico/Ithaka Electronic-Archiving Initiative
- 12148 National Library of France
- 78319 Google
- 64269 Digital Curation Centre
- 67531 University of North Texas
- 62624 New York University
- 15230 Rutgers University
- 88435 Princeton University
- 61001 University of Chicago
- 78428 University of Washington
- 13038 World Intellectual Property Organization
- 20775 University of California San Diego
- 29114 University of California San Francisco
- 28722 University of California Berkeley
- 21198 University of California Los Angeles

ARK Resolution



DAM-LR

PIDs Management Issues

PID infrastructures come at a cost:

- Added layer of infrastructure must be managed
  - Guarantee PID uniqueness
  - Update the urid/url mapping when moving the object
- Resolver service must run with high availability
- Must be very sure that the PIDs can be handled by our archives also in the long term.
    - Outlive the http protocol
    - Support conveying responsibility for PID domains
- In the end it all depends on the commitment
  - to manage the PIDs (and associated metadata) as well as
  - guarantee access to the resource.

Ideas

- Associate PIDs with citation info (see ARK)
  - Develop applications that track the use of this citation metadata
  - Possibly require authentication for this
- We also need PIDs for collections
  - Research is often based on collections of resources
  - Embedding 1000 pids + citation info not practical
  - Consider creation of collection pids that point to other pids + special citation info.
  - Special service is needed.

The End

DAM-LR

PIDs and Versioning

Some Terminology

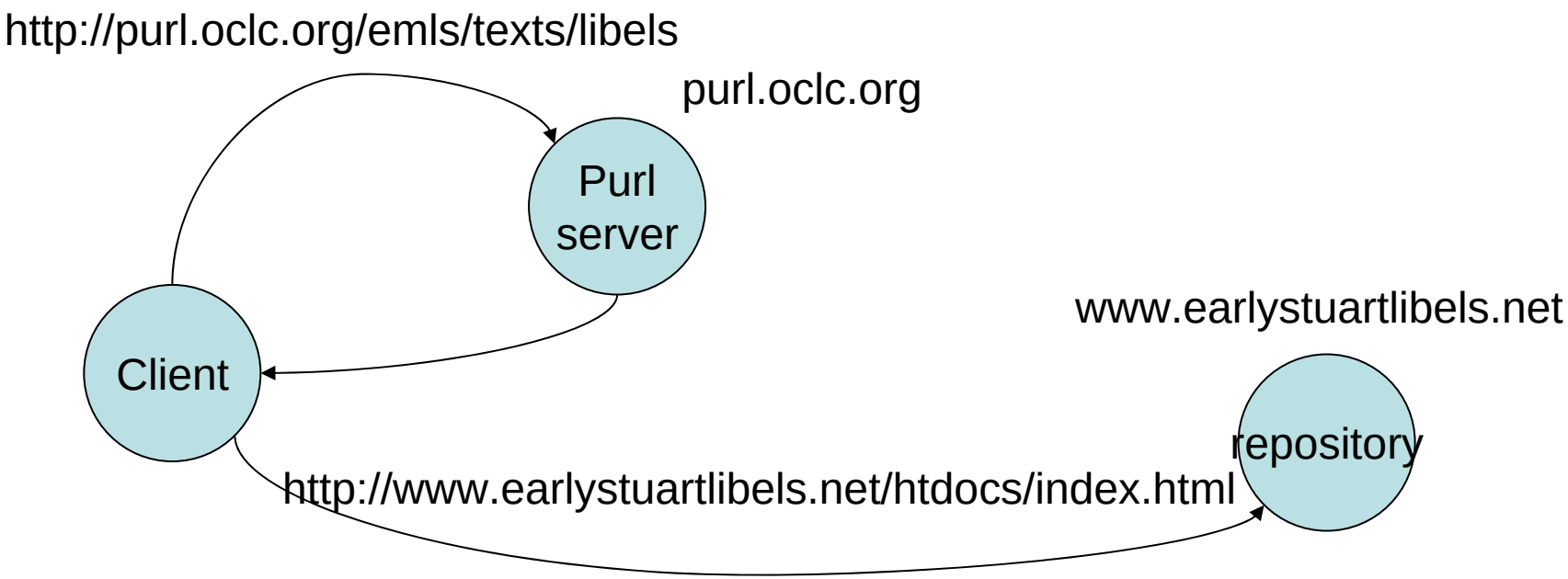
- (Digital) Identifier:
  - Name or label associated with an (electronic) resource. But …
- Namespace:
  - Domain or scope in which an identifier is created and is valid
- Persistence:
  - The identifier will permanently be associated with the resource and never be reused.
- Resolution:
  - The process of translates a resource identifier into the resource's location
- ARK
  - NAA: Name Assigning Authority
  - NAAN: Name Assigning Authority Number
  - NMAH: Name Mapping Authority Hostport

PURL

- Use of HTTP redirect

- Simple system
- direct actionable links
- Supported by many clients

- Protocol dependent
- Only one URL associated per purl
- No metadata resolving



Digital Object Identifier (DOI)

- Based on HS
- Created by publishers to:
  - facilitate electronic commerce,
  - enable copyright management
- INDECS metadata
- (approved) application profiles
- Heavily used comm. scient. publ., > 10M dois issued