

# Persistent Identification: The Handle System

Larry Lannom

Corporation for National Research Initiatives

<http://www.cnri.reston.va.us/>

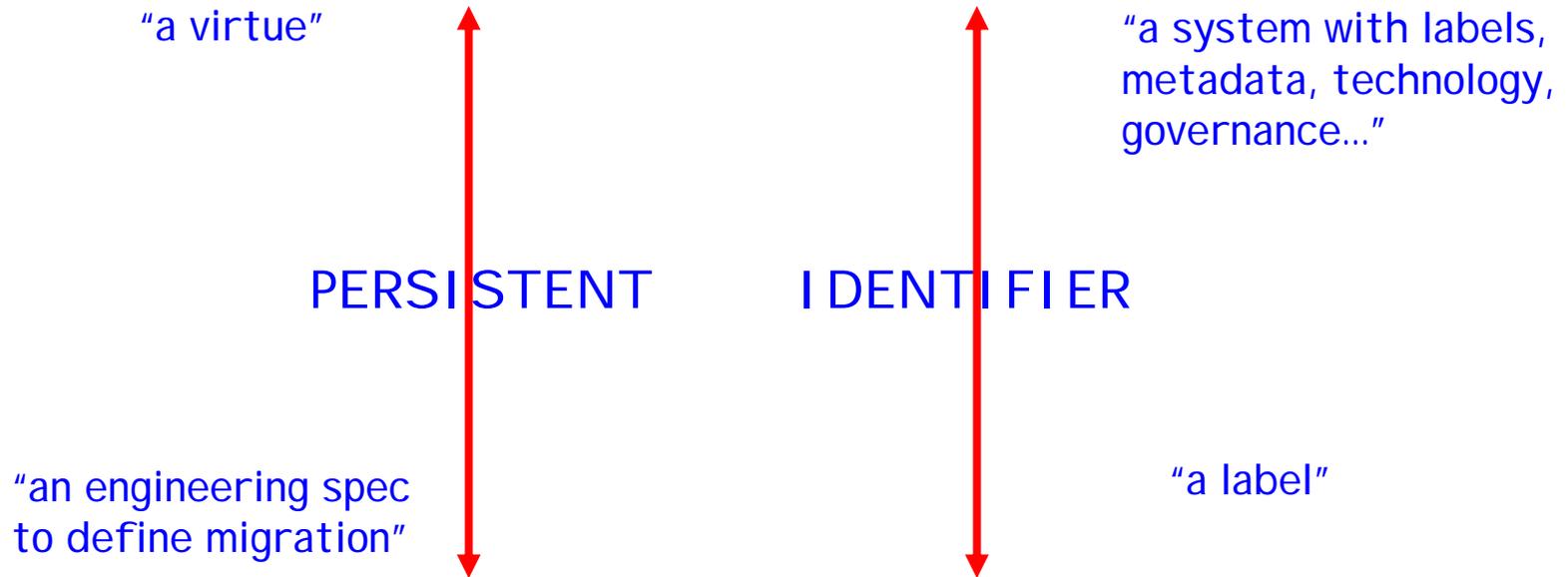
<http://www.handle.net/>

# The Handle System and Persistent Identification

- The evolving notion of Persistent Identifiers in the Library/Publishing/Document world
  - Acknowledgement - N. Paskin, Erpanet PI conf 2004
- Handle System overview
- Relationships

# The word trap...

---



- There are several meanings for "persistent" and "identifier" , so:
  1. Even if using only one word:
    - do you and I mean the same thing when we say e.g. "identifier"...?
  2. Some combinations of the two are essentially meaningless
    - category mistake ("the personality of a banana" )
- Philosophers solve this problem by "defining what functions you mean by this word?" (functional decomposition); but...

# I identifiers

---

- We all know our own back yard (“We all know what we mean”)
- Q: Why do we want persistent identifiers?
- A: For interoperability
- “persistence is interoperability with the future”
- We know what we mean, but others may not.
  - I identifiers assigned in one context may be encountered, and may be re-used, in another place (or time) - without consulting the assigner. You can't assume that your assumptions will be known to someone else.  
Interoperability = the possibility of use in services outside the direct control of the issuing assigner
- Interoperability is accelerated through automation:
  - Two key events:
  - 1966: automation of supply chains (I SBN)
  - 1994: automation of sharing resources (WWW)
- Increasing interoperability = increasing chance of breakdown

# Persistence

---

- "It is intended that the lifetime of a [persistent identifier] be permanent. That is, the [persistent identifier] will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name."
- [Persistent Identifier] = URN in IETF RFC 1737: Functional Requirements for Uniform Resource Names. (<http://www.ietf.org/rfc/rfc1737.txt>)

Technical and social infrastructure issues

# Persistence?

JISC Information Environment Architecture Standards Framework Version 1.1 May 2004



## 3. Web standards and file formats

This section outlines some broad Web guidelines with which all JISC IE Web sites should comply. In this document, the phrase 'JISC IE Web sites' refers to all Web sites associated with JISC IE service components.

JISC IE Web sites **must** be delivered using [HTTP 1.1](#) [4].

JISC IE Web sites should be accessible to all. All sites **must** achieve level A compliance with the The World Wide Web Consortium (W3C) [Web Accessibility Initiative Recommendations \(WAI\)](#) [5]. All sites **should** also achieve level AA compliance. This will ensure a high degree of usability for people with disabilities. Web sites **should** be accessible to a wide range of browsers and hardware devices (e.g. PDAs as well as PCs). Sites **should** be usable by browsers that support W3C recommendations such as [HTML/XHTML](#) [6], [Cascading Stylesheets \(CSS\)](#) [7] and [Document Object Model \(DOM\)](#) [8].

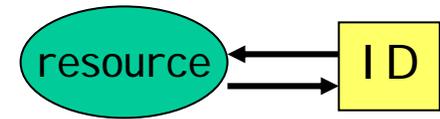
This document currently makes no specific recommendations about the file formats that should be used for various resource types (text, images, sounds, etc.). Such recommendations are made in the [Standards and Guidelines to Build a National Resource](#) [9] document (though it should be noted that this document is currently undergoing revision). However, sites **should** make use of open or de-facto standards whenever possible.

Every significant item that is made available through a JISC IE network service **should** be assigned a [URI](#) [10] that is reasonably persistent. This means that item URIs **should not** be expected to break for a period of 10-15 years after they have first been used. For this reason, JISC IE service components **should not** hardcode file format, server technology, service organisational structure or other information that is likely to change over a 10-15 year period into item URIs. If items become unavailable during that period, then the URI **should** resolve to a Web page that explains why the item is no longer available and what actions the end-user can take to obtain a copy of the item or similar resources. Furthermore, item URIs **should not** contain end-user-specific information, i.e. all item URIs should work for all end-users (albeit allowing for appropriate authentication challenges to be inserted into the process by which the URI is resolved).

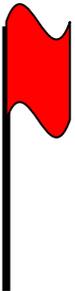
Resources that comprise a collection of items that are packaged together for management or exchange purposes **should** be packaged using the [IMS Content Packaging Specification](#) [11] if they are 'learning objects' (i.e. resources are primarily intended for use in a learning and teaching context and that have a specific pedagogic aim) or the [Metadata Encoding & Transmission Standard \(METS\)](#) [12].

# Two principles for persistent identification

---



1. *Obvious:* Assign ID to resource
  - Once assigned the number must identify the same resource
  - Beyond the lifetime of the resource, or the assigner
2. *Less obvious:* Assign Resource to ID
  - The resource must be "identified"
  - Must ensure it is always the same thing (bound)
  - Describe the resource "content" [with precision]
  - Failure to do this will ultimately break interoperability



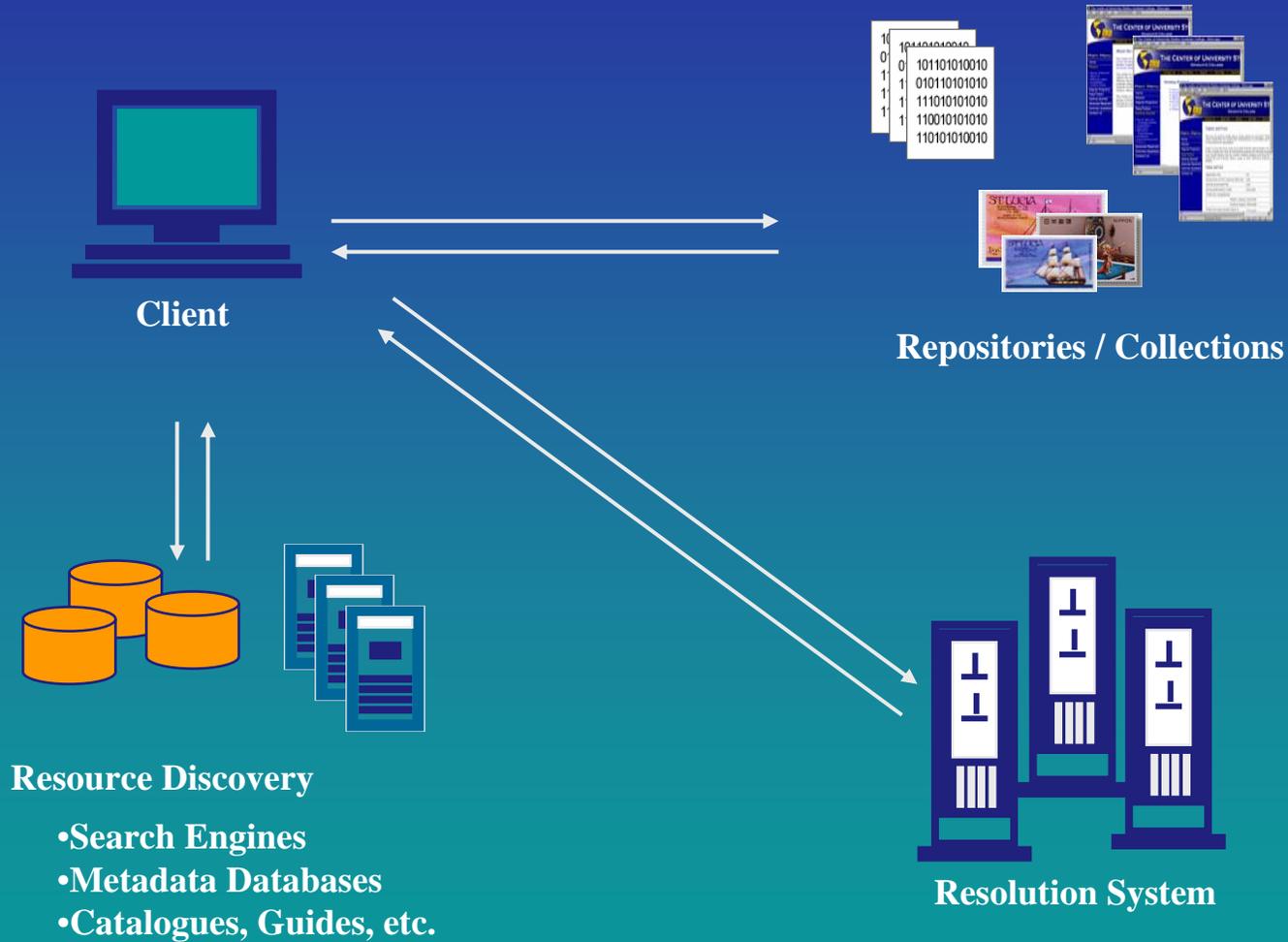
How far do we go in each? Depends on what we think is "good enough"

- Technologists have focussed on (1) [and "bags of bits/data structures"].
- The content/rights world (2) [and focus on "intellectual content"]
- Both viewpoints valid
- (2) is now becoming more relevant

# Digital Object Architecture - Goals

- Framework for managing Digital (Information) Objects
- Give it a name and talk to it
  - Don't worry about where it is
  - Don't worry about what it's made of
- Rise above details of application versions and content formats

# Digital Object Architecture

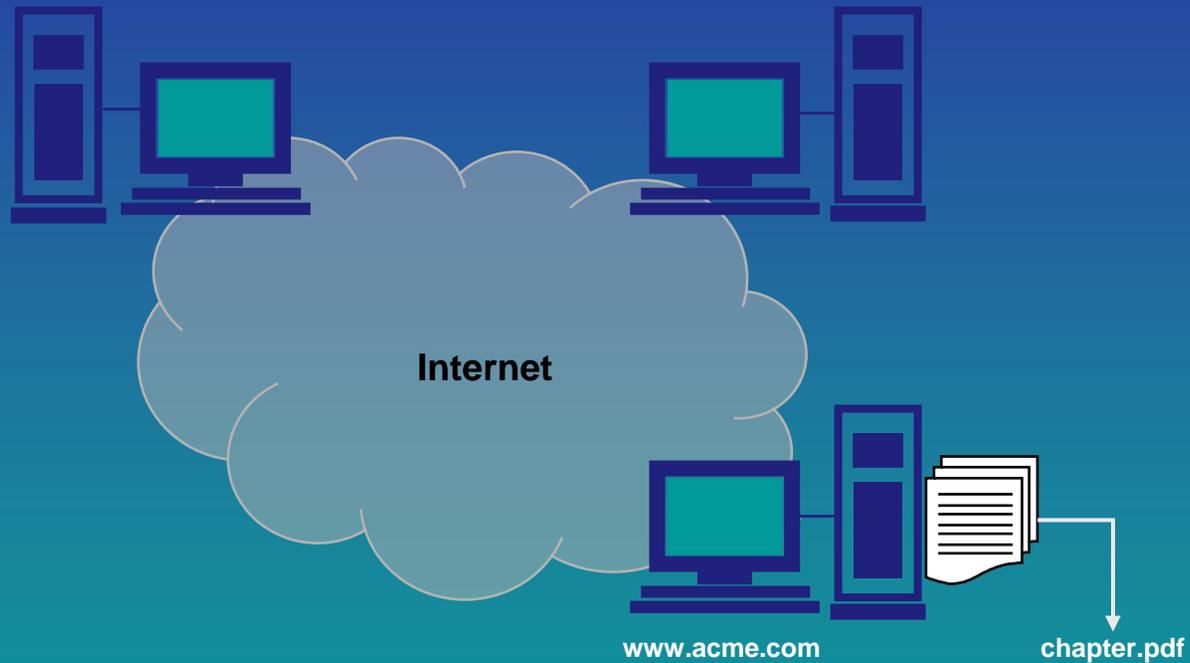


# Digital Object Architecture Components Handle System

- Go from name to attributes
- Fundamental indirection system for Digital Object management on the net
- No free lunch
  - Added layer of infrastructure
  - Must be managed

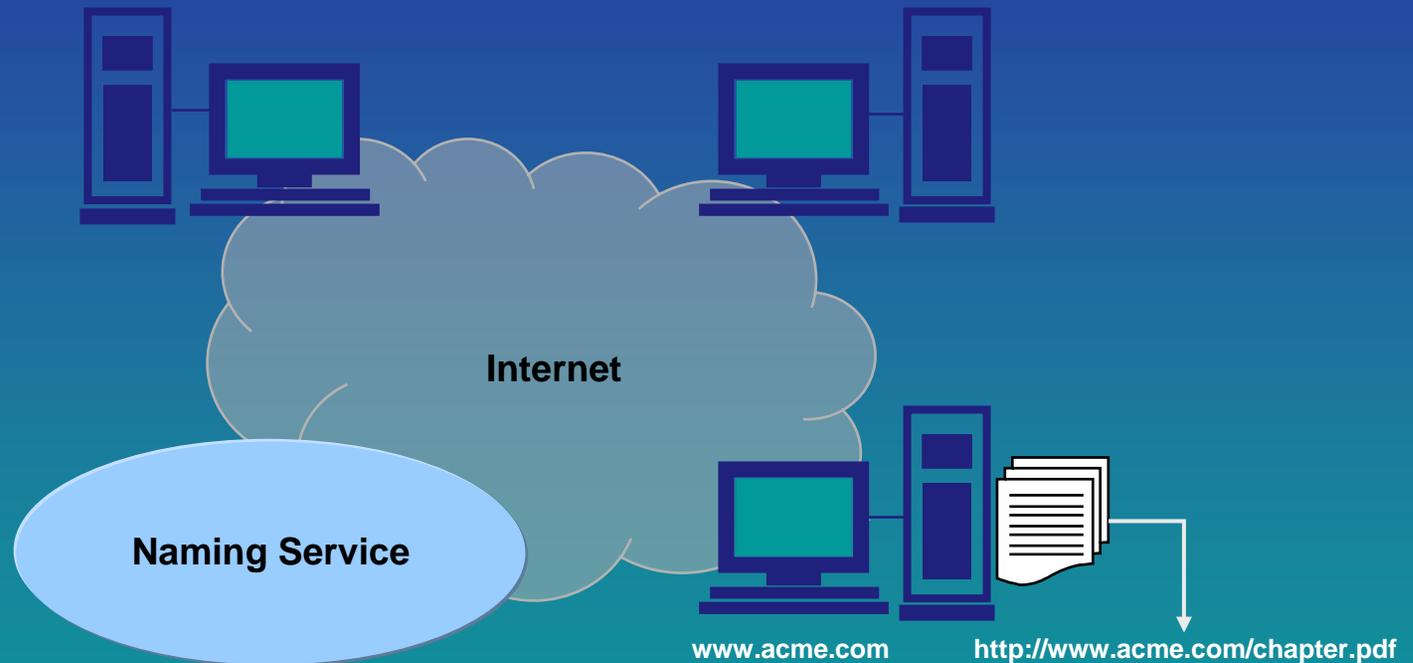
# Naming Resources on the Net

## The Problem



# Naming Resources on the Net

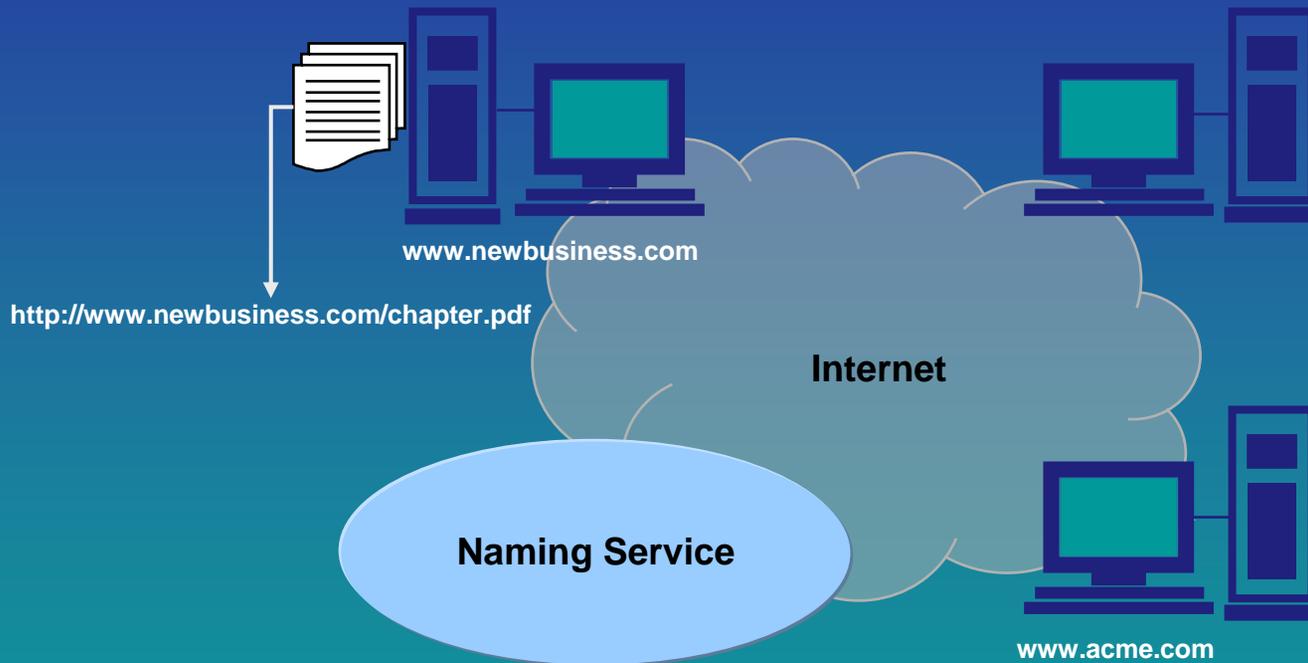
## The Solution



Name = Value(s)  
10.123/xyz = http://www.acme.com/chapter.pdf

# Naming Resources on the Net

## The Solution



Name = Value(s)

10.123/xyz = http://www.newbusiness.com/chapter.pdf

# CNRI Handle System

- Distributed, scalable, secure
- Enforces unique names
- Enables association of one or more typed values, e.g., URL, with each name
- Optimized for speed and reliability
- Open, well-defined protocol and data model
- Provides infrastructure for application domains, e.g., digital libraries, electronic publishing ...

# Handle System Usage

- Library of Congress
- DTIC (Defense Technical Information Center)
- IDF (International DOI Foundation)
  - CrossRef (scholarly journal consortium)
  - Enpia (Korean content management technology firm)
  - CDI (U.S. content management technology firm)
  - LON (U.S. learning object technology firm)
  - CAL (Copyright Agency Ltd - Australia)
  - TSO (U.K. publisher & info mgmt service provider)
  - MEDRA (Multilingual European DOI Registration Agency)
  - Nielsen BookData (bibliographic data - ISBN)
  - R.R. Bowker (bibliographic data - ISBN)
  - Office of Publications of the European Community
- NTIS (National Technical Information Service)
- DSpace (MIT + HP)
- CORDRA (ADL's Federated Content Repository Model)
- Various digital library production and research projects

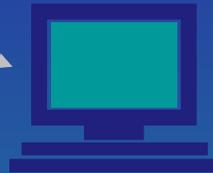
# Handles Resolve to Typed Data

Handle	Data type	Index	Handle data
10.123/456	URL	1	http://acme.com/...
	URL	2	http://a-books.com/...
	DLS	9	acme/repository
	HS_ADMIN	100	acme.admin/jsmith
	XYZ	12	1001110011110



# Handle Clients

Request to Client:  
Resolve hdl:10.1000/1



Client

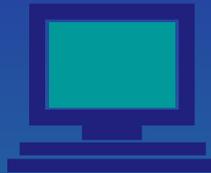
1. Sends request to Global to resolve 0.NA/10.1000 (naming authority handle for 10.1000)



Global Handle Registry

# Handle Clients

Request to Client:  
Resolve hdl:10.1000/1



Client

2. Global Responds with  
Service Information for 10.1000



Global Handle  
Registry

xcccXV	xC	xC	xC	...
xcccXV	xC	xC	xC	..
xccX	xC	xC	xC	..
xccX	xC	xC	xC	..
xcccXV	xC	xC	xC	..
xccX	xC	xC	xC	..
xccX	xC	xC	xC	..
xcccXV	xC	xC	xC	..
xccX	xC	xC	xC	..
xccX	xC	xC	xC	..

Service Information  
Acme Local Handle Service

# Handle Clients

XCCCXV	XC	XC	XC	...
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..

	IP Address	Port #	Public Key	...
<b>Primary Site</b>				
Server 1	123.45.67.8	2641	K03RLQ...	...
Server 2	123.52.67.9	2641	5&M#FG...	...
<b>Secondary Site A</b>				
Server 1	321.54.678.12	2641	F^*JLS...	...
Server 2	321.54.678.14	2641	3E\$T%...	...
Server 3	762.34.1.1	2641	A2S4D...	...
<b>Secondary Site B</b>				
Server 1	123.45.67.4	2641	N0L8H7...	...

Service Information - Acme Local Handle Service

# Handle Clients

XCCCXV	XC	XC	XC	...
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..

	IP Address	Port #	Public Key	...
<b>Primary Site</b>				
Server 1	123.45.67.8	2641	K03RLQ...	...
Server 2	123.52.67.9	2641	5&M#FG...	...
<b>Secondary Site A</b>				
Server 1	321.54.678.12	2641	F^*JLS...	...
Server 2	321.54.678.14	2641	3E\$T%...	...
Server 3	762.34.1.1	2641	A2S4D...	...
<b>Secondary Site B</b>				
Server 1	123.45.67.4	2641	N0L8H7...	...

Service Information - Acme Local Handle Service

# Handle Clients

XCCCXV	XC	XC	XC	...
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..
XCCCXV XCCX XCCX	XC XC XC	XC XC XC	XC XC XC	.. .. ..

	IP Address	Port #	Public Key	...
<b>Primary Site</b>				
Server 1	123.45.67.8	2641	K03RLQ...	...
Server 2	123.52.67.9	2641	5&M#FG...	...
<b>Secondary Site A</b>				
Server 1	321.54.678.12	2641	F^*JLS...	...
Server 2	321.54.678.14	2641	3E\$T%...	...
Server 3	762.34.1.1	2641	A2S4D...	...
<b>Secondary Site B</b>				
Server 1	123.45.67.4	2641	N0L8H7...	...

Service Information - Acme Local Handle Service

# Handle Clients

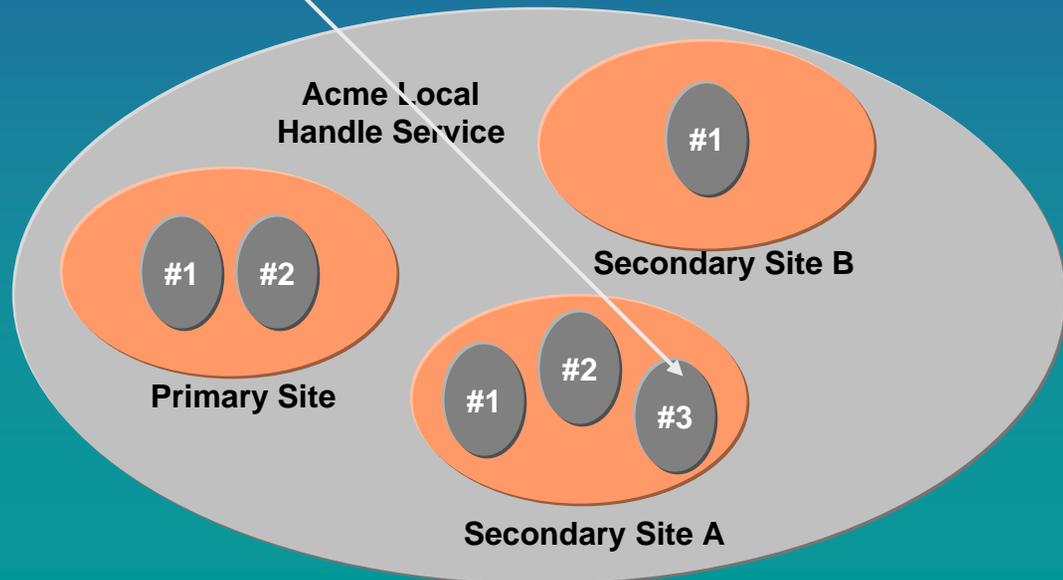
Request to Client:  
Resolve hdl:10.1000/1



Client

3. Client queries Server 3  
in Secondary Site A  
for 10.1000/1

Global Handle  
Registry



# Handle Clients

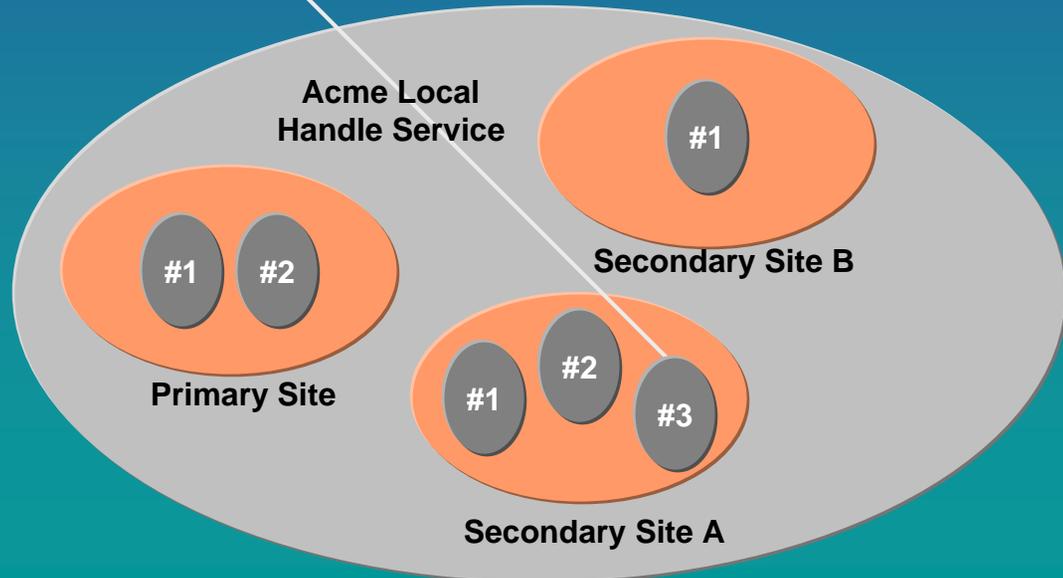
Request to Client:  
Resolve hdl:10.1000/1



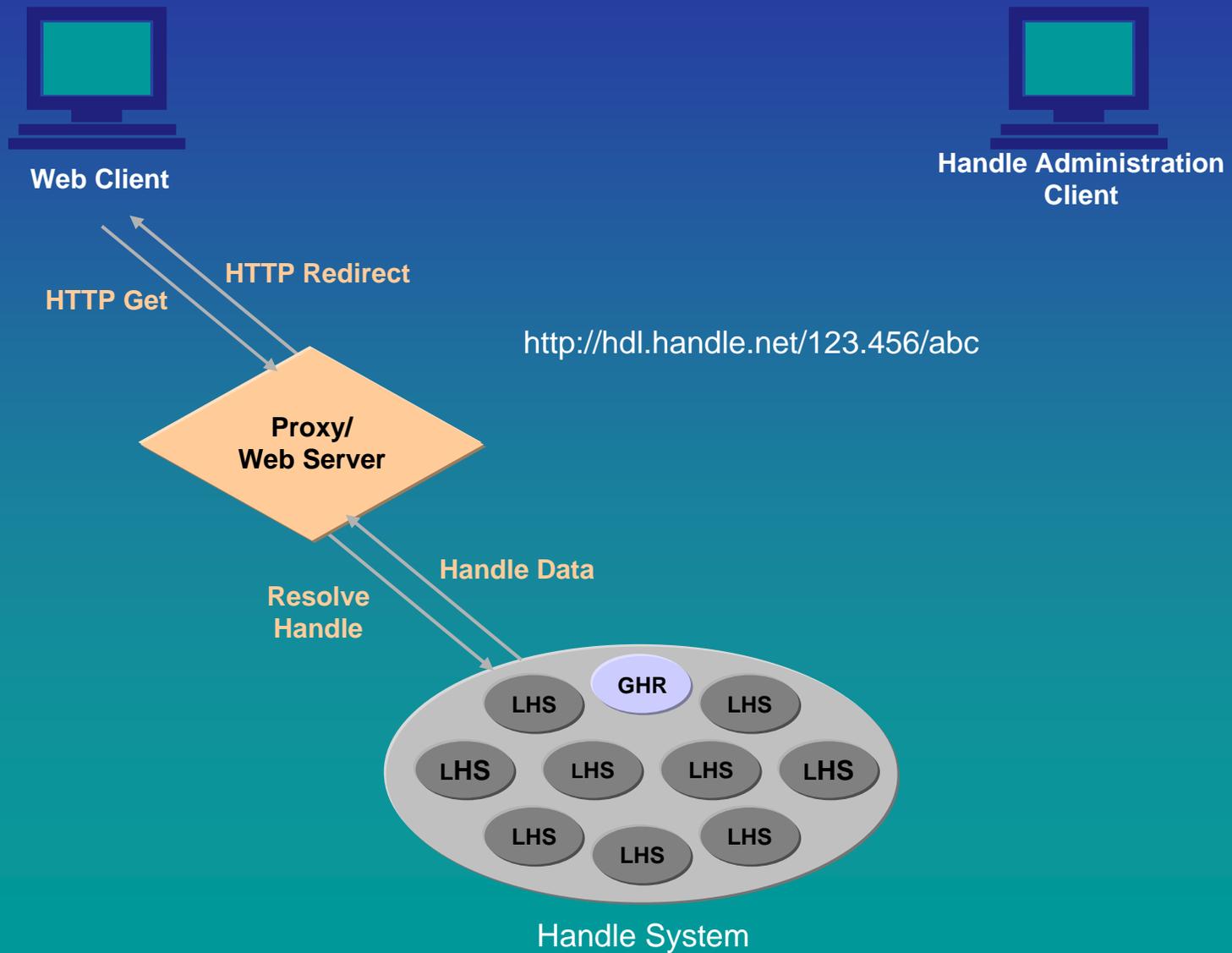
Client

Global Handle Registry

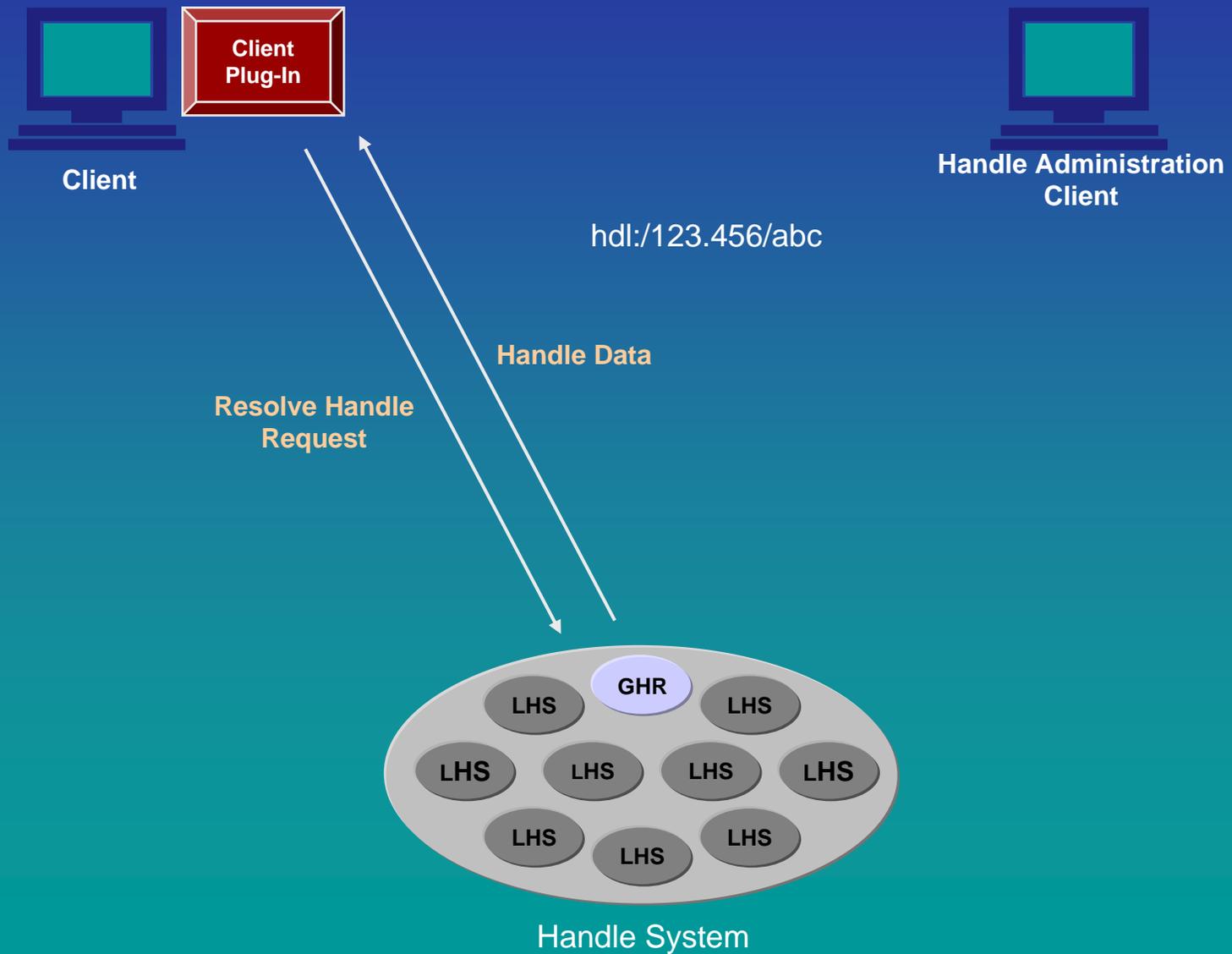
4. Server responds with  
handle data



# Handle Clients



# Handle Clients



# Handle Clients



Web



Handle Administration Client

HTTP

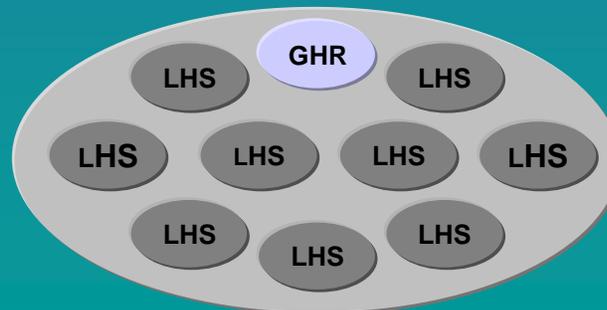


Web Server



Admin Forms

Handle Admin API



Handle System

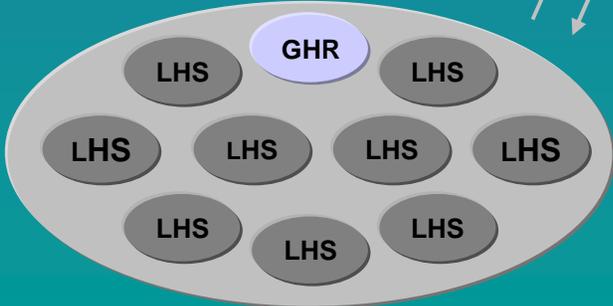
# Handle Clients



Web



Handle Administration Client



Handle System

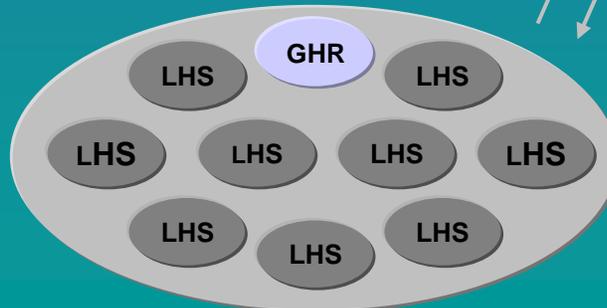


# Handle Clients



Web

*Handle Administration  
embedded in another  
process*



Handle System





# HS Administration

- Ownership is at the handle level
- Administrators defined by handles
- Administrator handles contain keys
- All admin transactions validated via challenge/response from server to client
- Allows distributed administration

# Handle System Usage

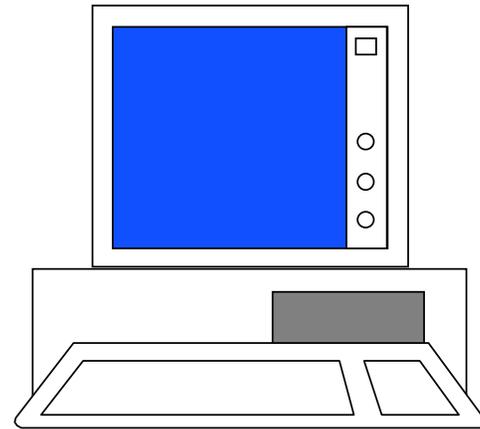
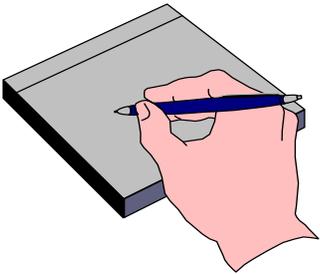
- Prefixes
  - DOI - 900
  - Other - 400
- Handles
  - DOI - 14M
  - Other - unknown
- Global
  - Three service sites (all currently in VA)
  - 10M resolutions last month

# Handle System Management and Standards

- Specification
  - RFC 3650: Overview
  - RFC 3651: Namespace and Service Definition
  - RFC 3652: Protocol
- HSAC - Handle System Advisory Committee
- URI/URL/URN
  - IETF votes for URN, we don't see any advantage
    - Extra layer of indirection, still need the native protocol
  - Many other groups pressuring for URI
  - What are the practical implications
  - Open to advice

# What are we identifying by this identifier?

---



Document on screen

Abstract work?

Manifestation of abstract work?

Version?

This HTML file?

All/some of these?

# Does it matter in everyday life?



Yes, it can do. e.g.:

## 1. Practical use of data. Example – journal article

- For the purpose of citation:
  - Count pdf, print, html as same
  - Citation refers to the abstract work (hence ISI, CrossRef)
- For the purpose of purchase:
  - Count pdf, print, html as different
  - Purchase refers to the manifestation
- Suppose I encounter a purchase system and try to use it for counting citations....
- Can I rely on a system now if I don't know what is being identified? Can others rely on the system long after I'm gone?

## 2. Legal implications: copyright

“My A is the same as your B and is my copyright...”

# Does A "mean the same as" B ?

---

- = in practice, does A need a different identifier from B?
  - versions; works and manifestations; editions
    - [e.g. two different e-book formats of the same work]
- For a machine, "A means same as B" = "A has same attributes as B"
- Which attributes? The answer is entirely contextual :
  - "Is A the same as B *for the purposes of ...?*"
  - = Do A and B belong to the same class for the purposes of ...
- For a machine, "for the purpose of" = "class having this set of attributes"
- We group similar things together; what is identified is usually a class
  - e.g. *the class of all copies of the hardback printed second edition of this book from this publisher* = the same ISBN
  - The class is defined by a set of attributes (metadata) (RDF, etc)
- No one thing is the same as another thing (or they wouldn't be two things)
  - "Roughly speaking, to say of two things that they are identical is nonsense, and to say of one thing that it is identical with itself is to say nothing at all." (L.W.)
  - Leibniz's Law (no two objects have exactly the same properties)
- Philosophy? philosophy = logic = automation

# Issues and themes for persistent identifier applications

---

## *ISSUES*

- What are we identifying with this identifier? [content not just bits]
- What are we resolving to from this identifier?
- What, if any, explicit metadata are we making available?
- How will the cost of providing the infrastructure be met ?

## *THEMES*

- Identification of entities of all forms
  - *To be used in variety of contexts*
- Appropriate use of metadata at appropriate level
  - *Development of ontology tools to describe entity relationships*

[llannom@cnri.reston.va.us](mailto:llannom@cnri.reston.va.us)

[www.handle.net](http://www.handle.net)