Identifiers

CS431 – Architecture of Web Information Systems
Carl Lagoze – Cornell University – Feb. 7 2007

BEWARE

A Lecture with lots of questions and not as many answers!

Acknowledgments

- Stuart Weibel – OCLC
- Herbert Van de Sompel – LANL
- Andy Powell – EduServ
- Norman Paskin – International DOI Foundation

Identifiers

- Provide a key or *handle* linking abstract concepts to physical or perceptible entities
- Provide us with a necessary figment of persistence
- They are perhaps the one *essential* and common form of *metadata*
- Why bother?
  - Finding things
  - Comparing things
  - Referring to things (Citations)
  - Asserting ownership over things

Identity <-> Change <-> Persistence

- Paradox: reality contains things that persist and change over time
  - Heraclitus and Plato: can you step into the same river twice?
  - Ship of Theseus: over the years, the Athenians replaced each plank in the original ship of Theseus as it decayed, thereby keeping it in good repair. Eventually, there was not a single plank left of the original ship. So, did the Athenians still have one and the same ship that used to belong to Theseus

Identity <-> Change <-> Persistence

I have lots of identifiers

- Carl Jay Lagoze, Dad, Hey you
- 123-456-7890 (SSN)
- 1234-5678-1234-1234 (Visa Card)
- FZBMLH (US Airways locator on January 18 flight to San Diego)

What do we want from identifiers?

- Global uniqueness
- Authority
- Reliability
- Appropriate functionality
  - Resolution
  - Other services
- Persistence

Identifier Issues

- Object granularity
- Identifier Context
  - Object atomicity
  - Part/whole relationships
- Location independence
  - Multiple location resolution
- Human vs. machine generation and resolution
- Administration (centralized vs. decentralized)
- Intrinsic semantics

Opaque versus Semantic Identifiers

- DOI:10.1045/3451/13x.4
- http://store.apple.com/1-800-MY-APPLE/WebObjects/AppleStore

- Should identifiers carry semantics?
  - People like semantic identifiers
  - Semantic Drift can be a problem
    - Words and names change meaning over time
  - Semantics can compromise persistence
    - Organizations/People/Concepts change over time
  - Semantics is culturally laden

Varieties of semantics

- Opaque
  - Nothing can be inferred, including sequence
  - Cannot be reverse-engineered (feature or bug?)
- Low-resolution date semantics
  - LCCN 99-087253
- Encoded semantics
  - ISBN 1-58080-046-7
  - Country codes… agency codes… checksums…
- Sequential Semantics
  - OCLC numbers
- Name/Word Semantics
  - Work Name/Chapter Name/Section Name

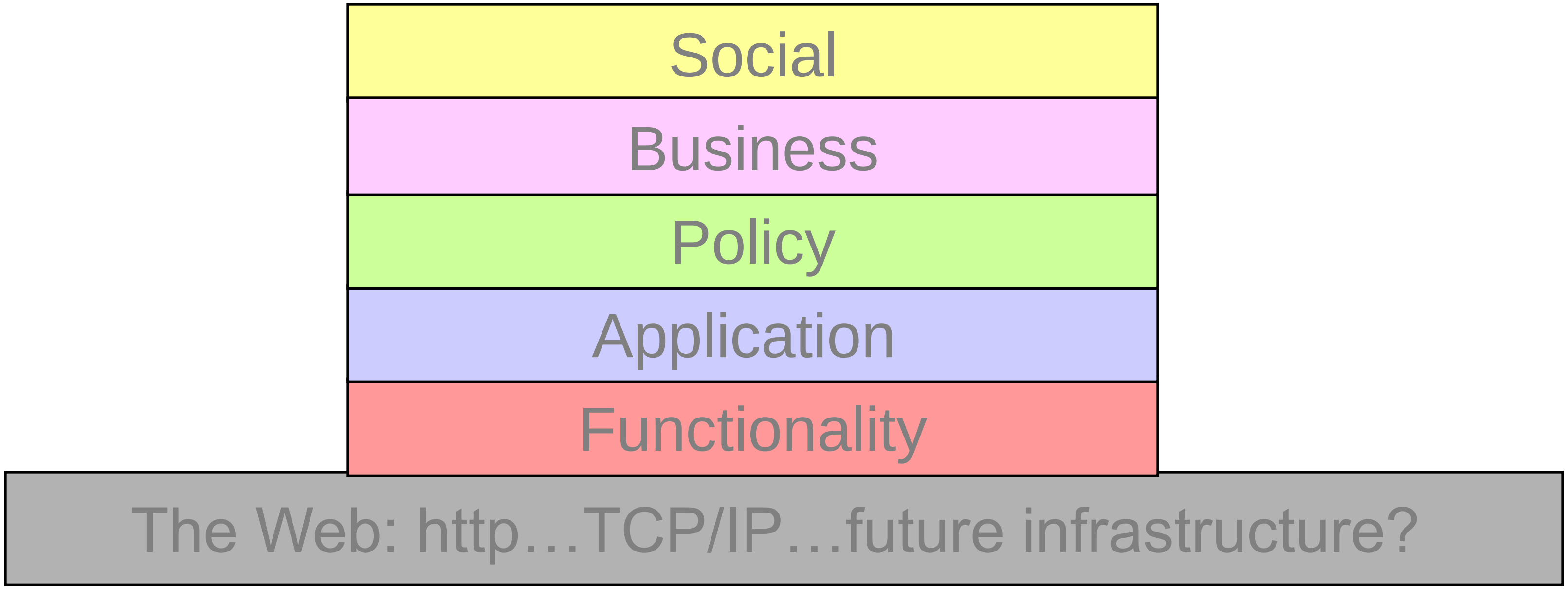
Lots of (non-digital) Identifier Standards

- ISBN (International Standard Book Number)
  - Origin 1966 U.K.
  - ISO 2108 1970
  - Uniquely identifies each edition and variation of a book
  - Number is semantically meaningful (components)
    - prefix/country code/pub code/item #/checksum
  - International administration (>150 countries)
- ISSN (International Standard Serial Number)
  - Uniquely identifies every serial (not issue or volume)
  - Semantically meaningless (anonymous)
  - International administration
- Lots of others
  - Recording Code, Tech Report, Audiovisual

http://www.collectionscanada.ca/iso/tc46sc9/index.htm

The Identifier Layer Cake

- Identifiers come in many sizes, flavors, and colors… what questions do we ask?



Social

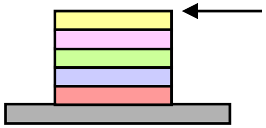
Business

Policy

Application

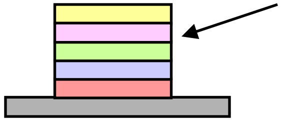
Functionality

The Web: http…TCP/IP…future infrastructure?



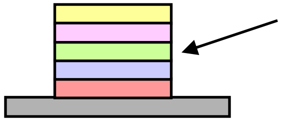
Social Layer

- The only guarantee of the usefulness and persistence of identifier systems is the commitment of the organizations which assign, manage, and resolve identifiers
- Whom do you trust?
  - Governments?
  - NGOs?
  - Cultural heritage institutions?
  - Commercial entities?
  - Non-profit consortia?
- We trust different agencies for different purposes at different times



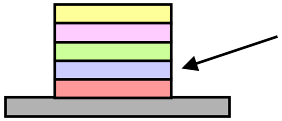
Business layer

- Who pays the cost?
- How, and how much?
- Who decides (see governance model)?



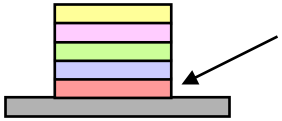
Policy Layer

- Who has the ‘right’ to assign or distribute Identifiers?
- Who has the ‘right’ to resolve them or offer serves against them?
- What are appropriate assets for which identifiers can be assigned, and at what granularity?
- Can identifiers be recycled?
- Can ID-Asset bindings be changed?



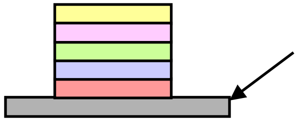
Application Layer

- What underlying dependencies are assumed?
  - http… tcp/ip…(bar code|RFID) scanners…
- What is the nature of the systems that support assignment, maintenance, resolution of identifiers?
- Are servers centralized? federated? peer to peer?
- How is uniqueness assured?



Functional Layer:
Operational characteristics of Identifiers

- Is it globally unique? (easy)
- How does it ‘behave’? What applications recognize it and act on it appropriately?
- Do identifiers need to be matched to the characteristics of the assets they identify?
- Do humans need to read and transcribe them?

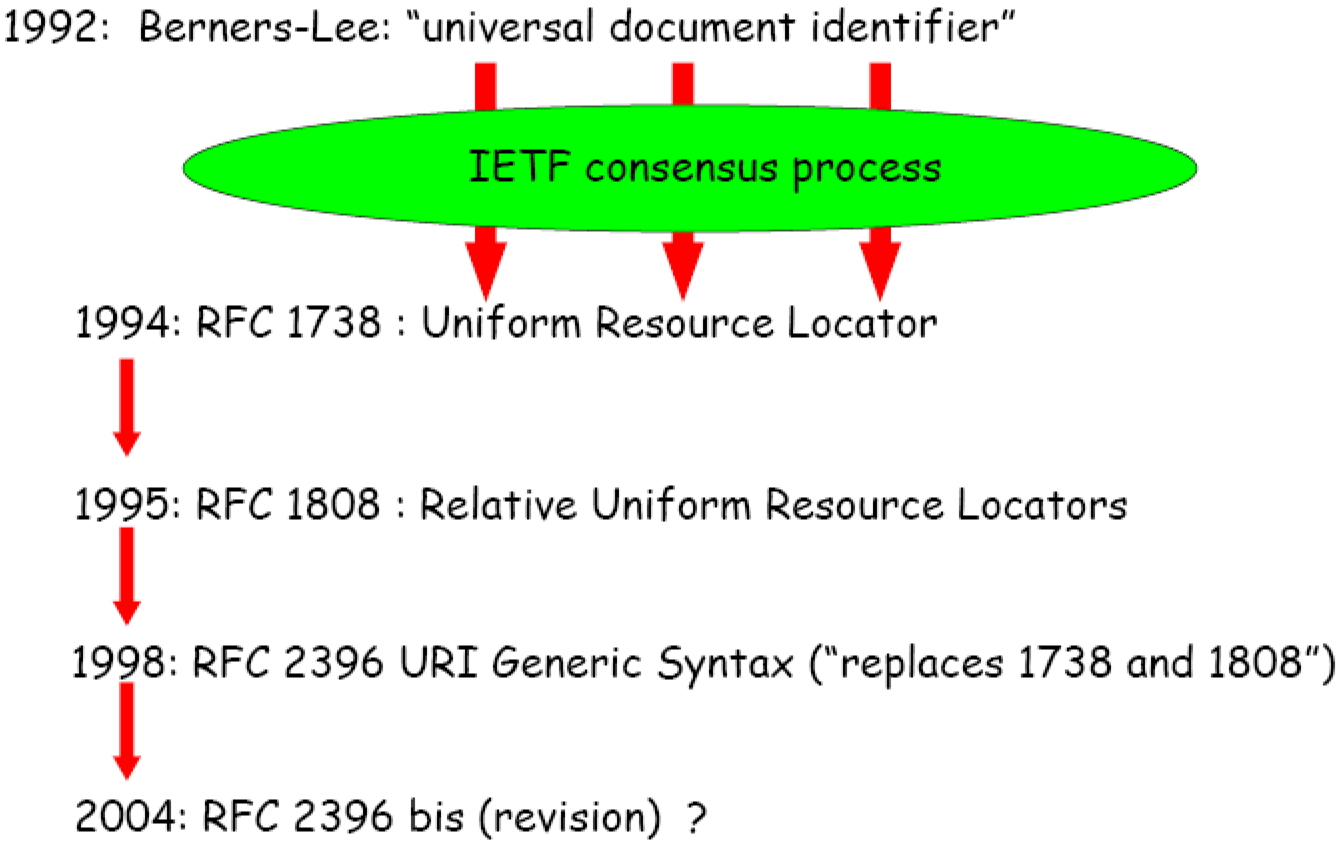


Technology layer: The Web

Some fundamental questions:

- Must our identifiers be URIs?
- Must they be universally actionable?
- If so, what is the desired action?
- Is there ever a reason to use a URI other than an http-URI as an identifier?

Persistent identifiers on the web



1992: Berners-Lee: “universal document identifier”

IETF consensus process

1994: RFC 1738 : Uniform Resource Locator

1995: RFC 1808 : Relative Uniform Resource Locators

1998: RFC 2396 URI Generic Syntax (“replaces 1738 and 1808”)

2004: RFC 2396 bis (revision) ?

Why isn’t DNS sufficient (parenthetical comment)

- Issue of semantic vs. non-semantic names
- Changing ownership
- Hierarchical legacy of DNS is sometimes inappropriate

Pure Identifiers versus pure Locators

- But *locators* and *identifiers* are not the same…or are they?
- In Web-space, they are close:
  - Not every *identifier* is a *locator*, but every *locator* is an *identifier*

- And do we need identifiers when Google-like search makes "identifier-free" location possible?

Robust Hyperlinks

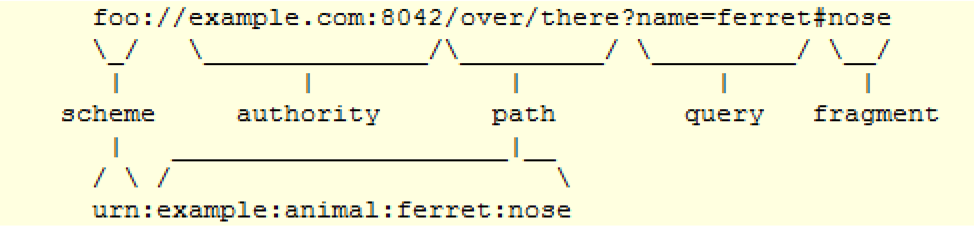
- http://www.dlib.org/dlib/july00/wilensky/07wilensky.html
- Characteristic of document (metadata) is computed automatically via fingerprint of its content.
- "Lexical" signatures: The top n words of a document chosen for rarity, subject to heuristic filters to aid robustness.
  - "a TF-IDF-like" measure
  - Five or so words are sufficient
- Can be used to locate document (via search engine) after it is moved

Robust Hyperlinks – Why does this work?

- Number of terms on Web is reportedly close to 10,000,000.
- If terms were distributed independently, the probability of 5 even moderately common terms occurring in more than one document is very small.
  - In fact, picking 3 terms restricted to those occurring in 100,000 documents works pretty well.
  - Many documents contain very infrequently used words.
  - There is lots of room for independence to be off, and to play with term selection for robustness, etc..

URI: Universal Resource Identifier

- Generic *syntax* for identifiers of resources
- Defined by RFC 2396
- Syntax: <scheme>:<scheme-specific-part>
  - ftp://ftp.is.co.za/rfc/rfc1808.txt
  - http://www.ietf.org/rfc/rfc2396.txt
  - mailto:John.Doe@example.com
  - urn:oasis:names:specification:docbook:dtd:xml:4.1.2
- Hierarchically-organized, components in order of decreasing significance



```
  foo://example.com:8042/over/there?name=ferret#nose
  \_/   \______________/\_________/ \_________/ \__/
   |           |            |            |        |
scheme     authority       path        query   fragment
   |   _____________________|__
  / \ /                        \
  urn:example:animal:ferret:nose
```

URI Schemes (as of 2005 06 03) http://www.iana.org/assignments/uri-schemes

ftp	File Transfer Protocol
http	Hypertext Transfer Protocol
gopher	The Gopher Protocol
mailto	Electronic mail address
news	USENET news
nntp	USENET news using NNTP access
telnet	Reference to interactive sessions
wais	Wide Area Information
prospero	Prospero Directory
z39.50s	Z39.50
z39.50r	Z39.50 Retrieval
cid	content identifier
mid	message identifier
vemmi	versatile multimedia
Interfaceservice	service location
imap	internet message access protocol
nfs	network file system protocol
acap	application configuration access
protocolrtsp	real time streaming protocol
tip	Transaction Internet Protocol
pop	Post Office Protocol v3
data	data
dav	dav
opaquelocktoken	opaquelocktoken
sip	session initiation protocol
sips	secure session intitiaion protocol
tel	telephone
fax	fax

modem	modem
ldap	Lightweight Directory Access Protocol
https	Hypertext Transfer Protocol Secure
soap.beep	soap.beep
soap.beeps	soap.beeps
xmlrpc.beep	xmlrpc.beeps
xmlrpc.beeps	xmlrpc.beeps
urn	Uniform Resource Names
go	go
h323	H.323
ipp	Internet Printing Protocol
tftp	Trivial File Transfer Protocol
mupdate	Mailbox Update (MUPDATE) Protocol
pres	Presence
im	Instant Messaging
mtqp	Message Tracking Query Protocol
iris.beep	iris.beep
dict	dictionary service protocol
snmp	Simple Network Management Protocol
crid	TV-Anytime Content Reference Identifier
tag	tag

Reserved URI Scheme Names:

afs	Andrew File System global file names
tn3270	Interactive 3270 emulation sessions
mailserver	Access to data available from mail servers

Why is RFC 2396 so big?

- Character encodings
- Escaping Characters
- Partial and relative URIs
  - e.g. chap2/start.html, /top/next/part.html, #head1
  - Algorithms for establishing base URL and attaching relative reference to it
- URI Equivalence

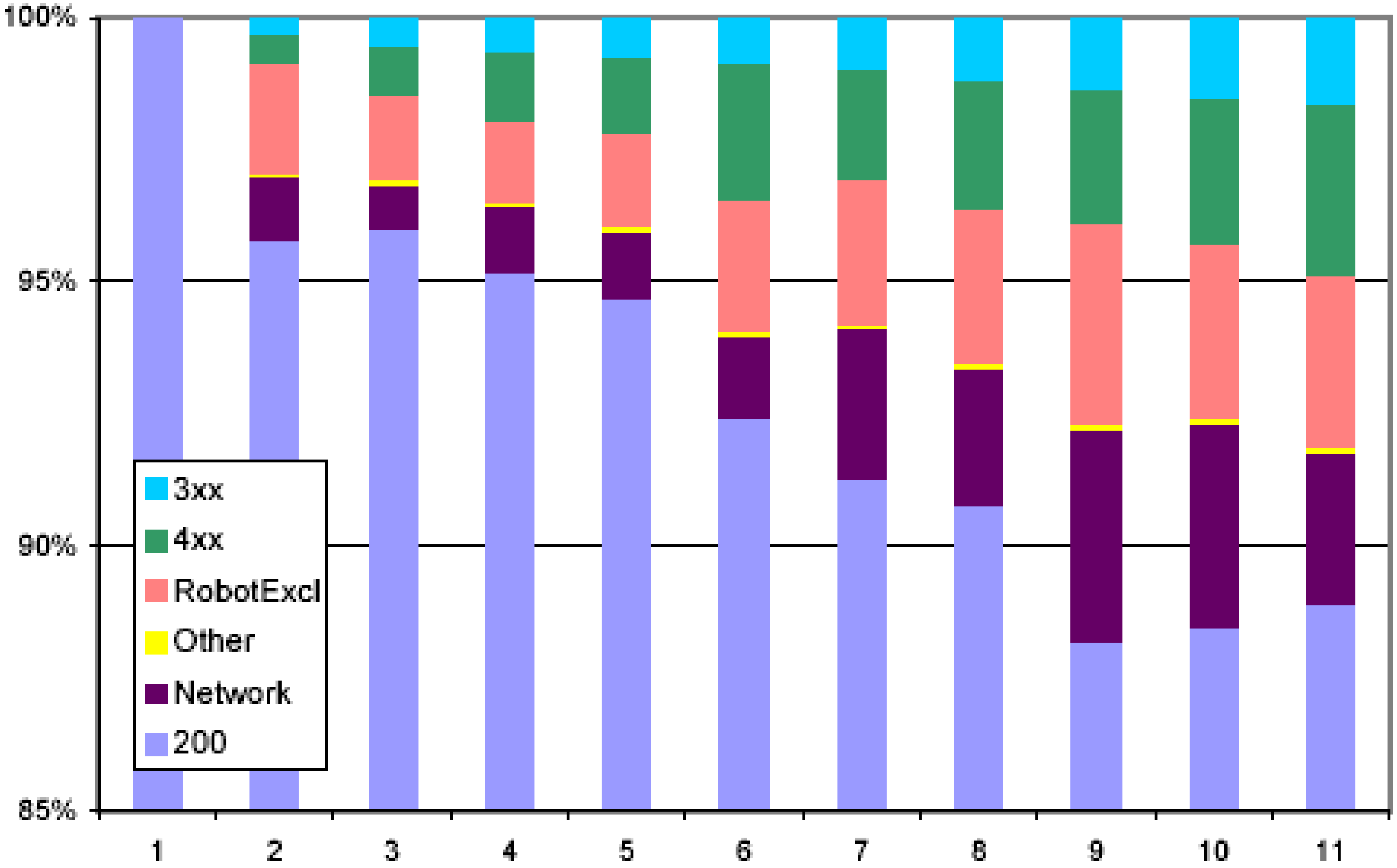
URL: Universal Resource Locator

- Deprecated term but we'll use it here
- String representation of the location for a resource that is available via the Internet
- Use URI syntax
- Scheme has function of defining the access (protocol) method. Used by client to determine the protocol to "speak".
  - http://an.org/index.html - open socket to an.org on port 80 and issue a GET for index.html
  - ftp://an.org/index.html - open socket to an.org on port 21, open ftp session, issue ftp get for index.html….

UR(I)L Issues

- Persistence
  - “link rot”
- Location dependence
- Valid only at the item level
  - What about works, expressions, manifestations
- Multiple resolution
  - “get the one that is cheapest, most reliable, most recent, most appropriate for my hardware, etc.”
- Non-digital resources?
- How about identifying representations?

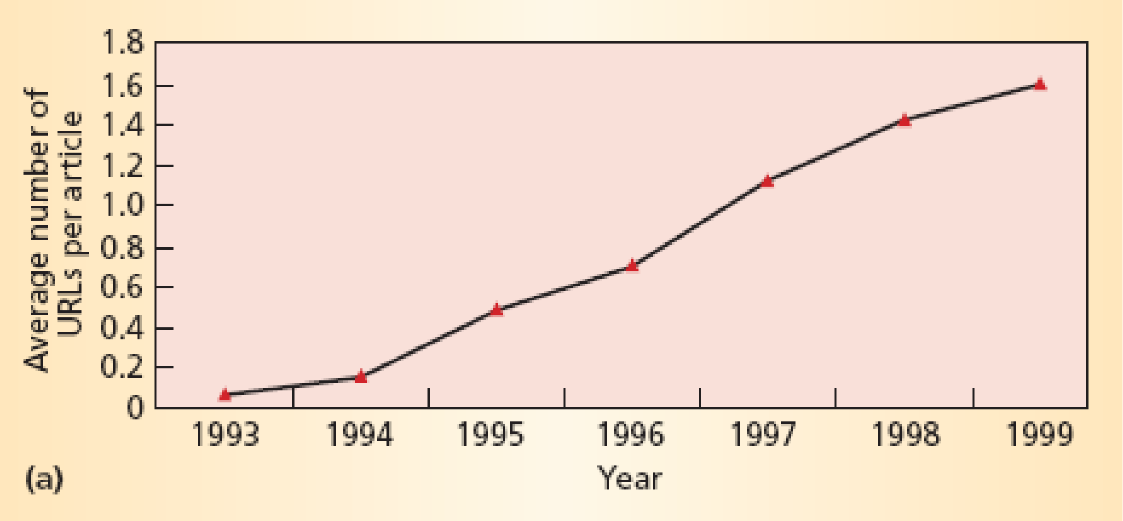
Link-rot

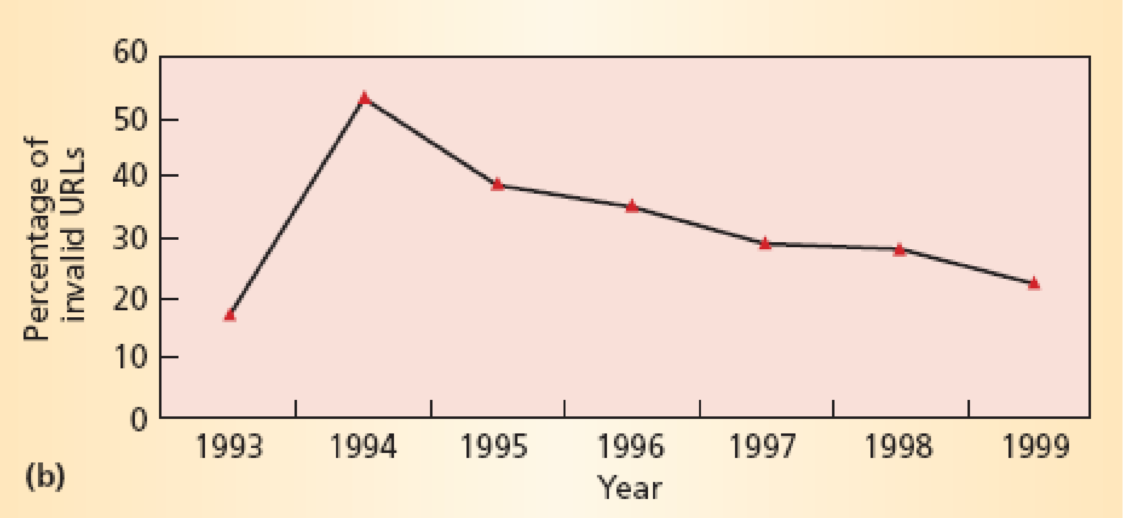


crawls ran consecutively, starting on 5 Dec. 2002 and ending on 12 Feb. 2003

http://www2003.org/cdrom/papers/refereed/p097/P97%20sources/p97-fetterly.html

Link-rot





http://ieeexplore.ieee.org/iel5/2/19496/00901164.pdf?isnumber=&arnumber=901164

The identifier persistence issue

“No scheme or syntax guarantees persistence of any kind”

John Kunze, California Digital Library

URI’s – The Web Gurus View
Henry Thompson W3C

- The web works because you can
  - View source
  - Follow your nose
  - Write URIs on the side of a bus
  - Use generic tools
  - Redirect, cache and proxy
- The Web is hands-down the most successful distributed name-based system the world has yet seen
  - Hmmm… Postal addresses, phone #’s?
- Ergo anyone designing a persistent identifier system should start from the assumption that http URIs are sufficient for their *technology* needs.

Arguments for http URI’s

- Application Ubiquity: every Web application recognizes them. Achieving similar ubiquity for other URI schemes is very difficult
- Relies on a well-proven distributed global lookup system (DNS)
  - Any naming system will have to have a lookup system
- Actionable identifiers are good – immediacy is a virtue
- If the Web is displaced, everyone has the problem of coping; if you invent your own solution, and it is displaced, you are isolated
- Using Non-ubiquitous identifiers will make it harder to maintain persistence over time by complicating the technical layer, which will compromise the ability to sustain long-term institutional commitments
- Focuses on non-technology issues involved in producing persistence

Cool URIs don’t change

Tim Berners-Lee 1998

http://www.w3.org/Provider/Style/URI

What makes a cool URI?

A cool URI is one which does not change.

What sorts of URI change?

URIs don't change: people change them

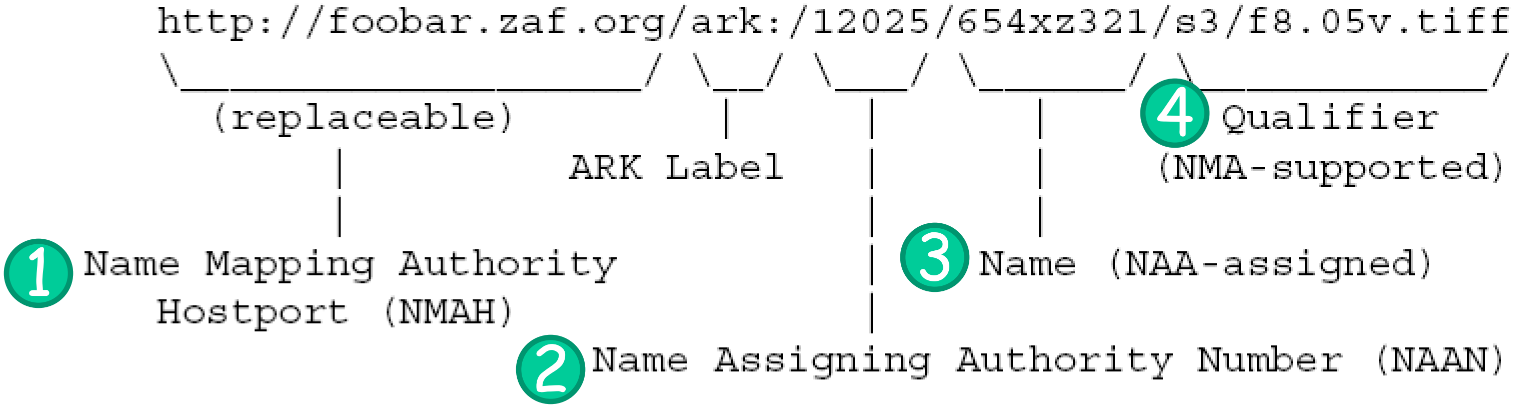
Archival Resource Key (ARK) Tools

Adding more "persistence" to URIs
by separating the provider of naming
services from the assigner of names

http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

ARK Summary

Instead of one Name Authority: Assigning Authority + Mapping Authorities



```
    http://foobar.zaf.org/ark:/12025/654xz321/s3/f8.05v.tiff
    \_____________________/ \__/ \___/ \______/ \____________/
       (replaceable)          |      |       |   4 Qualifier
             |          ARK Label    |       |    (NMA-supported)
             |                       |       |
1 Name Mapping Authority             |   3 Name (NAA-assigned)
    Hostport (NMAH)                  |
                    2 Name Assigning Authority Number (NAAN)
```

1 **= current service provider; identity inert; replaceable**
2 **= organization that originally assigned the id**
3 **= name originally assigned to the abstract object, often opaque**
4 **= extension disclosing object hierarchy & variants, often non-opaque**

ARK usage

Two ARKs accessing the same thing

http://loc.gov/ark:/12025/654xz321

http://rutgers.edu/ark:/12025/654xz321

Access to metadata -- add a '?'

http://loc.gov/ark:/12025/654xz321?

Access to support statement -- add '??'

http://loc.gov/ark:/12025/654xz321??

- 3 minimal requirements to be an ARK
  - An archive that can't do all 3 -- trustworthy?
  - Is an ARK persistent? Maybe. Have to *ask*.

OCLC's PURL

- PURL: Persistent Uniform Resource Locators
- They look like URLs… they ***ARE*** URLs
- No new technology, no new protocols, no plugins
- PURLs take advantage of inherent redirection facility in the HTTP protocol
- A simple toolset for managing names and namespaces

http://www.purl.org

PURL Syntax

- A PURL is a URL.

http://purl.oclc.org/OCLC/PURL/FAQ

protocol resolver address path (asset name)

PURL resolvers use standard http *redirects* (3xx status) to return the actual URL.

PURL Namespaces

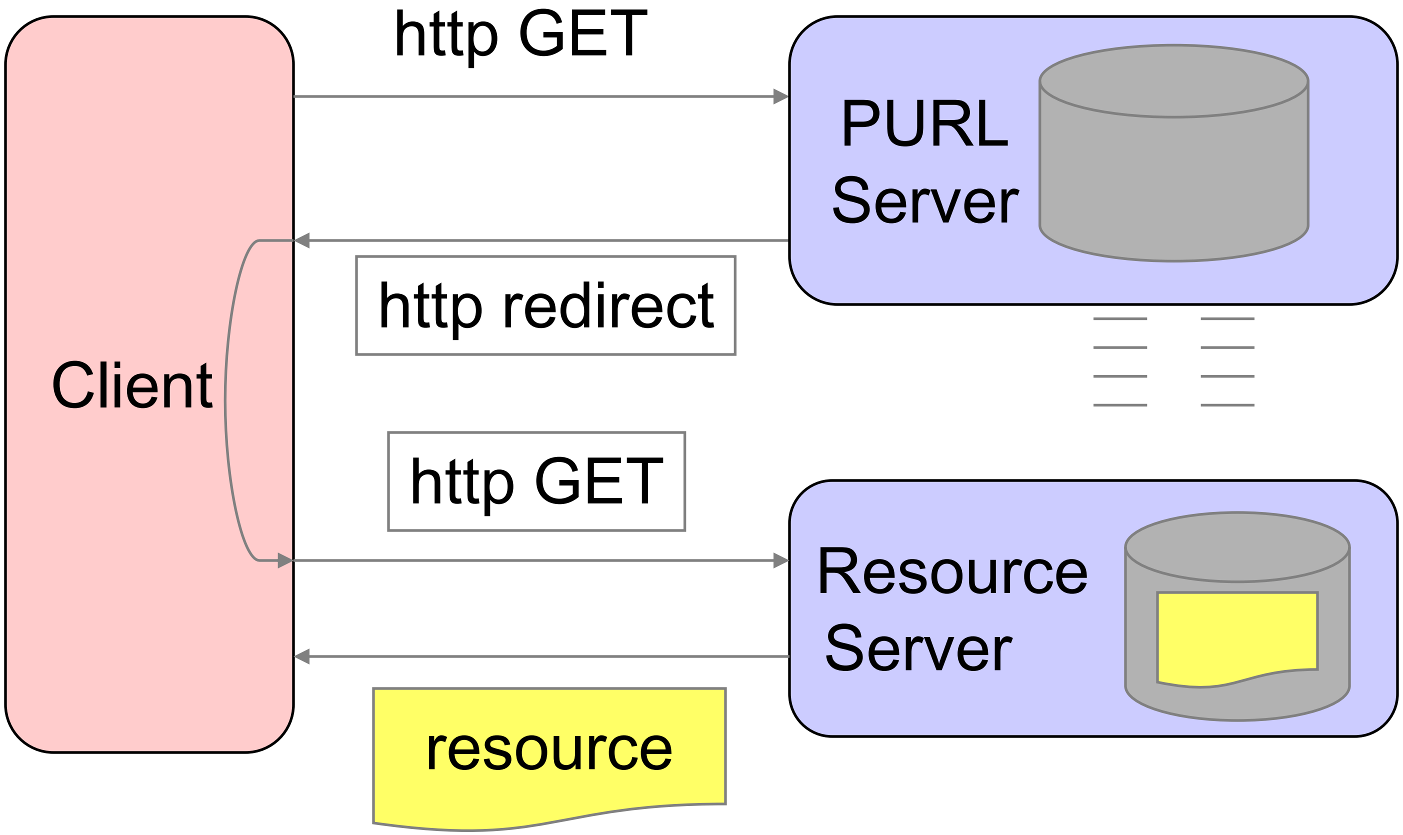
A PURL provides a local (not-global namespace)

http://purl.oclc.org/keith/home

is different from

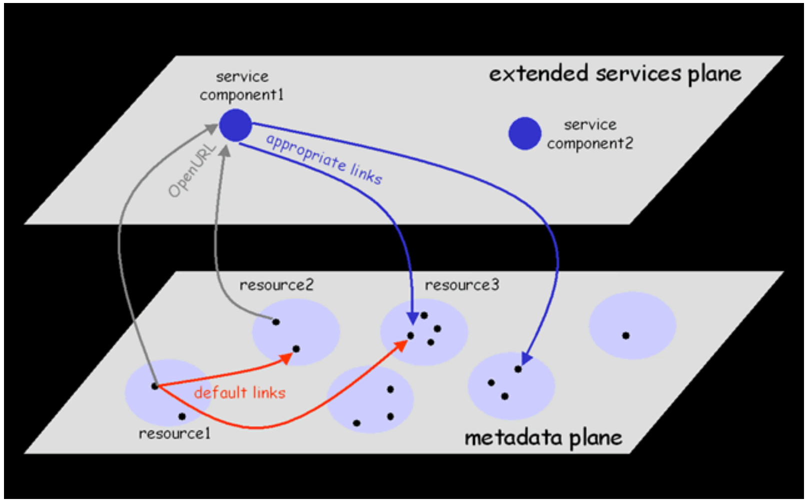
http://purl.stanford.edu/keith/home

OCLC PURL Resolution

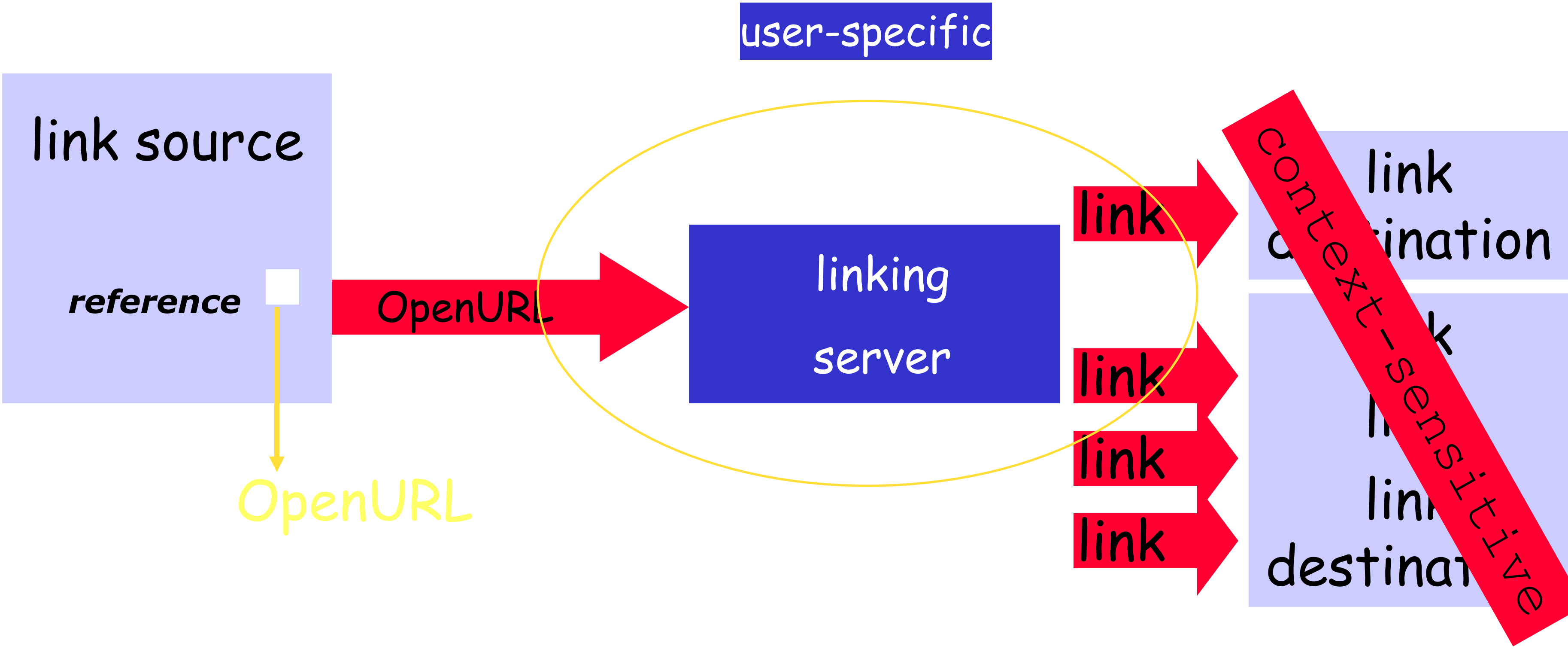


openURL: Making links context sensitive

- Why?
  - "Appropriate item" differs for each user
  - Licensing locality
  - Some users may want a choice (abstract, full text, etc.)
- Conceptualize link as service rather than object targeted.
- OpenURL
  - Transports metadata about the work to…
  - A localized service that interprets the metadata and provides contextualized choices to the user.



OpenURL linking

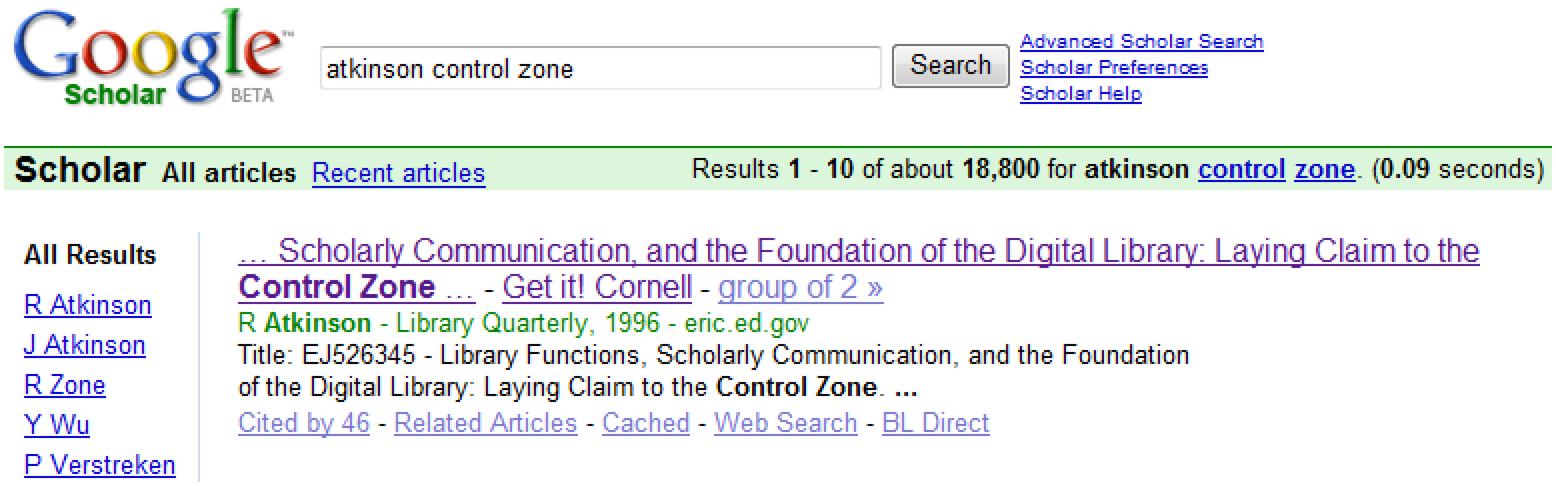


Components of an OpenURL

- Base-URL – Service component that accepts the openURL
- Object Description – Identifying information about an object (e.g., the identifier of a resource, metadata about the resource)
- Origin Description – Identifying information about origin of request.

http://www.ukoln.ac.uk/distributed-systems/openurl/

Google Scholar and OpenURL



Results **1** - **10** of about **18,800** for **atkinson control zone**. (**0.09** seconds)

... Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the **Control Zone** ... - Get it! Cornell - group of 2 »
R **Atkinson** - Library Quarterly, 1996 - eric.ed.gov
Title: EJ526345 - Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the **Control Zone**. **...**
Cited by 46 - Related Articles - Cached - Web Search - BL Direct

http://scholar.google.com/scholar?hl=en&lr=&q=atkinson+control+zone

The Silver Bullet: URN – Universal Resource Name

- “globally unique, persistent names”
- Independence from location and location methods

```
<URN> ::= "urn:" <NID> ":" <NSS>
```

- `NID`: namespace identifier
- `NSS`: namespace-specific string
- examples:
  - urn:ISSN:1234-5678
  - urn:isbn:9044107642
  - urn:doi:10.1000/140

Handles: Names for Internet Resources

- Naming system for location-independent, persistent names
- One name, multiple resolutions
- http://www.handle.net

The resource named by a Handle can be:

- A library item
- A collection of library items
- A catalog record
- A computer
- An e-mail address
- A public key for encryption
- etc., etc., etc. ....

Syntax of Handles

<naming_authority>/<locally_unique_string>

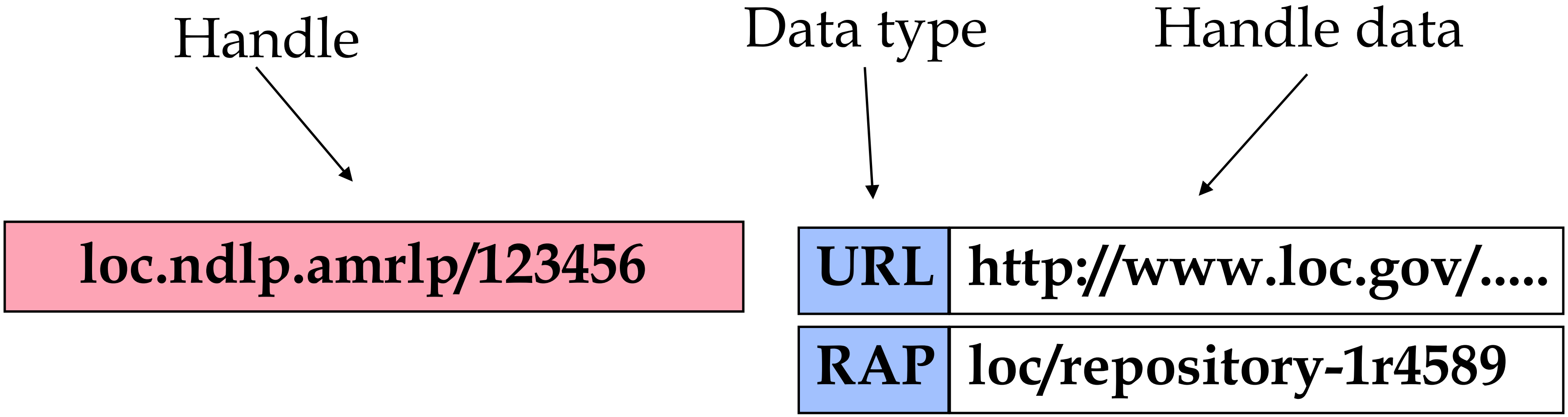
or

hdl:<naming_authority>/<locally_unique_string>

Examples

10.1234/1995.02.12.16.42.21;9	(date-time stamp)
cornell.cs/cstr-94.45	(mnemonic name)
loc/a43v-8940cgr	(random string)

Example of a Handle and its Data Used to Identify Two Locations



Handle

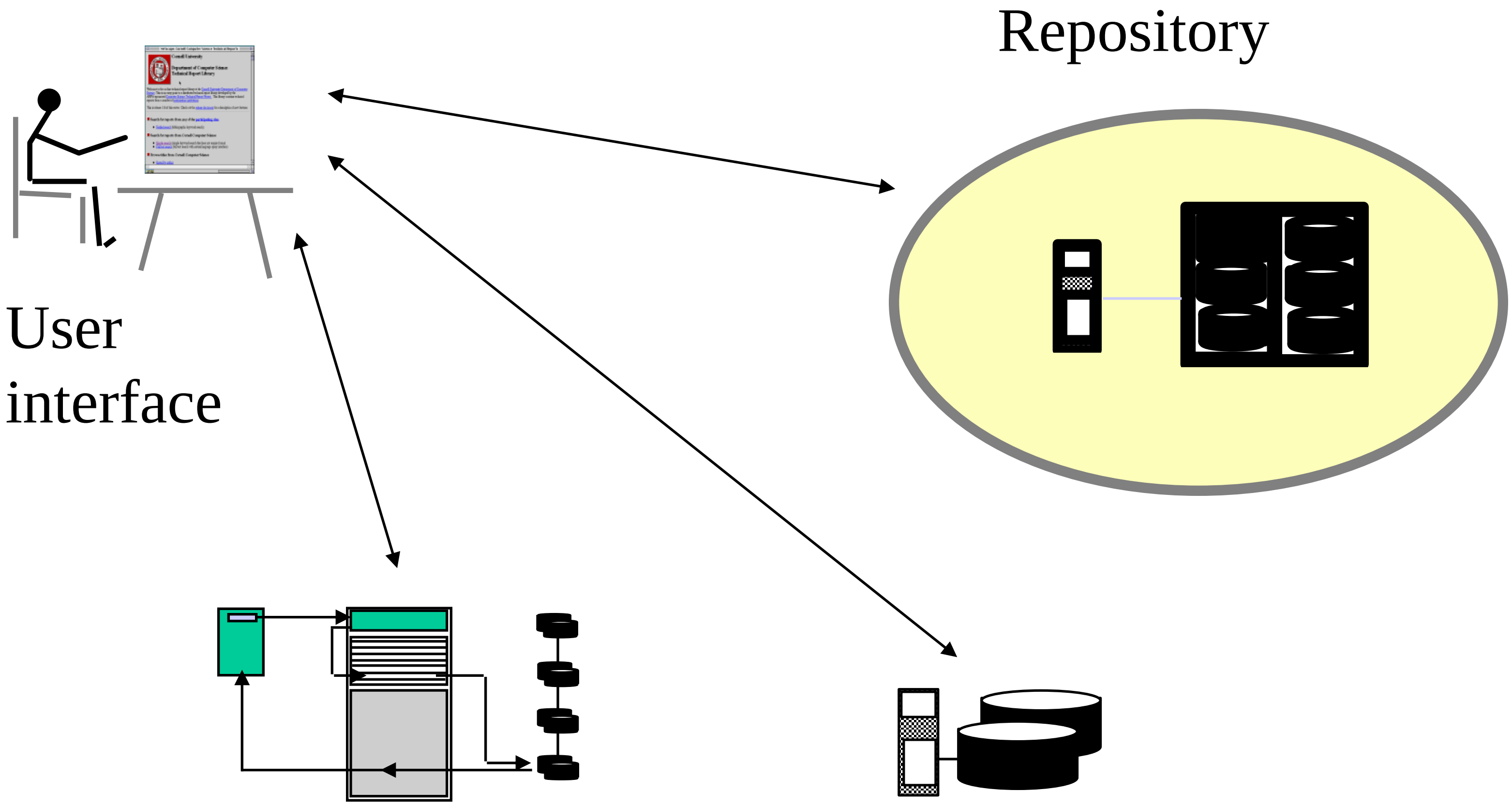
Data type

Handle data

loc.ndlp.amrlp/123456

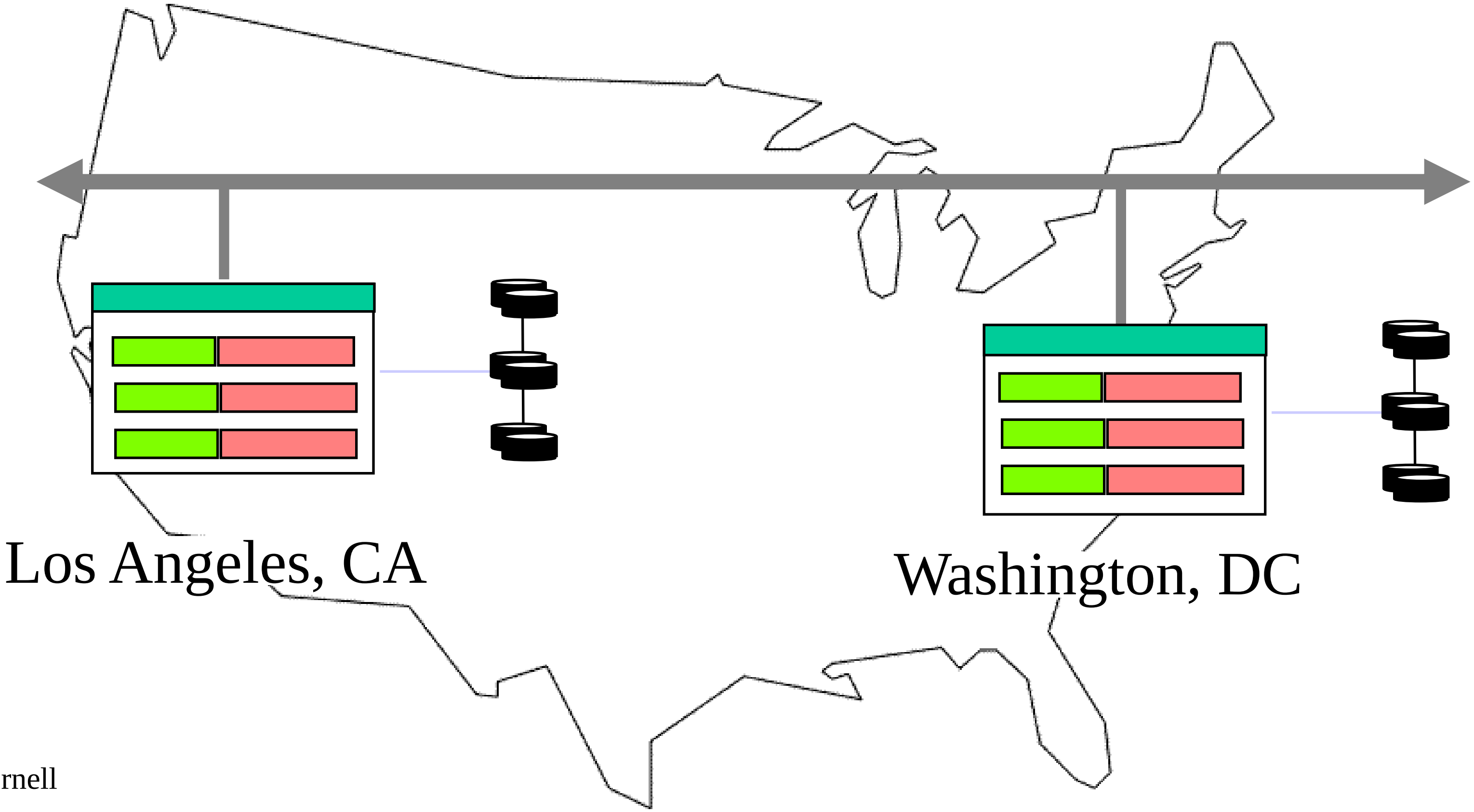
URL	http://www.loc.gov/.....
RAP	loc/repository-1r4589

Use of Handles in a Digital Library

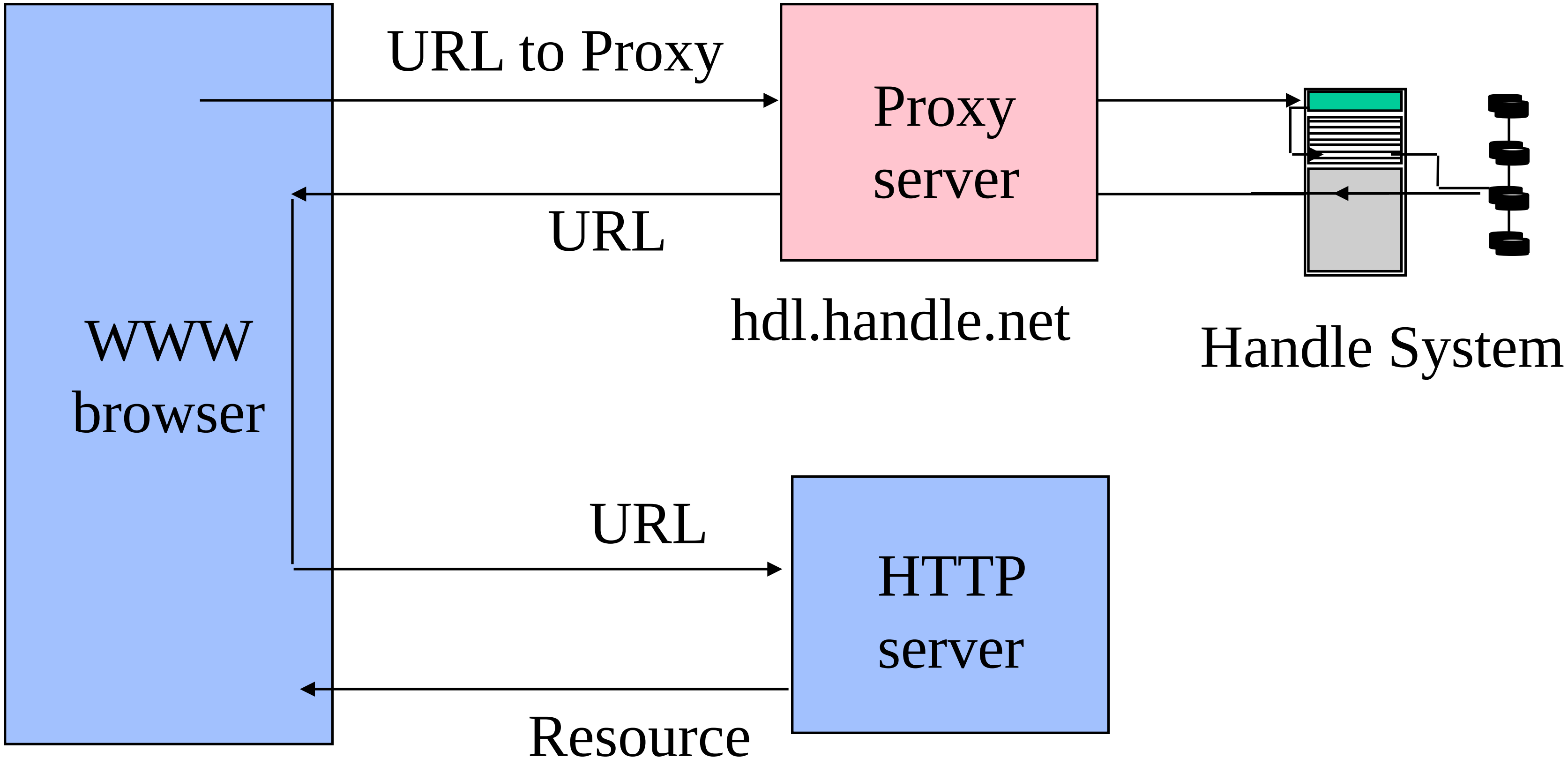


Replication for Performance and Reliability

Example: the Global Handle System



Proxy Resolution



DOI – Digital Object Identifier

- Technology and social infrastructure for naming
- Established by publishers for persistent naming of entities (articles, journals, conference proceedings)
- Cognizant of FRBR elements
- Underlying technology is handle system
  - "persistent" names
    - Persistence is fortified by social underpinnings
    - Rules for establishing registration agencies
  - Multiple resolution
  - Registration/mechanism has metadata associated with it
- doi:10.1000/186

Why haven’t URNs caught on beyond certain communities?

- Complexity of systems
- One size does not fit all - special purpose URN schemes have been successful, e.g., PubMed ID, Astrophysics BibCode
- No guarantee of persistence – longevity is an organizational not technical issue
- Requires well-regulated administrative systems
- Absence of “killer” applications – although reference linking is emerging

Conclusions

- There is no established “answer” the identification problem
  - Lots of identify wars
  - Turf protecting
- In reality there are different needs with different appropriate solutions
- URIs do work as an appropriate technological solution and must always be considered.