Identifiers

CS431 – Architecture of Web Information Systems
Carl Lagoze – Cornell University – Feb. 6 2006

Acknowledgments

- Stuart Weibel – OCLC
- Herbert Van de Sompel - LANL

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
- Human vs. machine generation and resolution
- Administration (centralized vs. decentralized)
- Intrinsic semantics
- Type specificity

Opaque versus Semantic Identifiers

- Should identifiers carry semantics?
  - People like semantic identifiers
  - Semantic Drift can be a problem
  - Semantics can compromise persistence
  - Semantics is culturally laden

Varieties of semantics

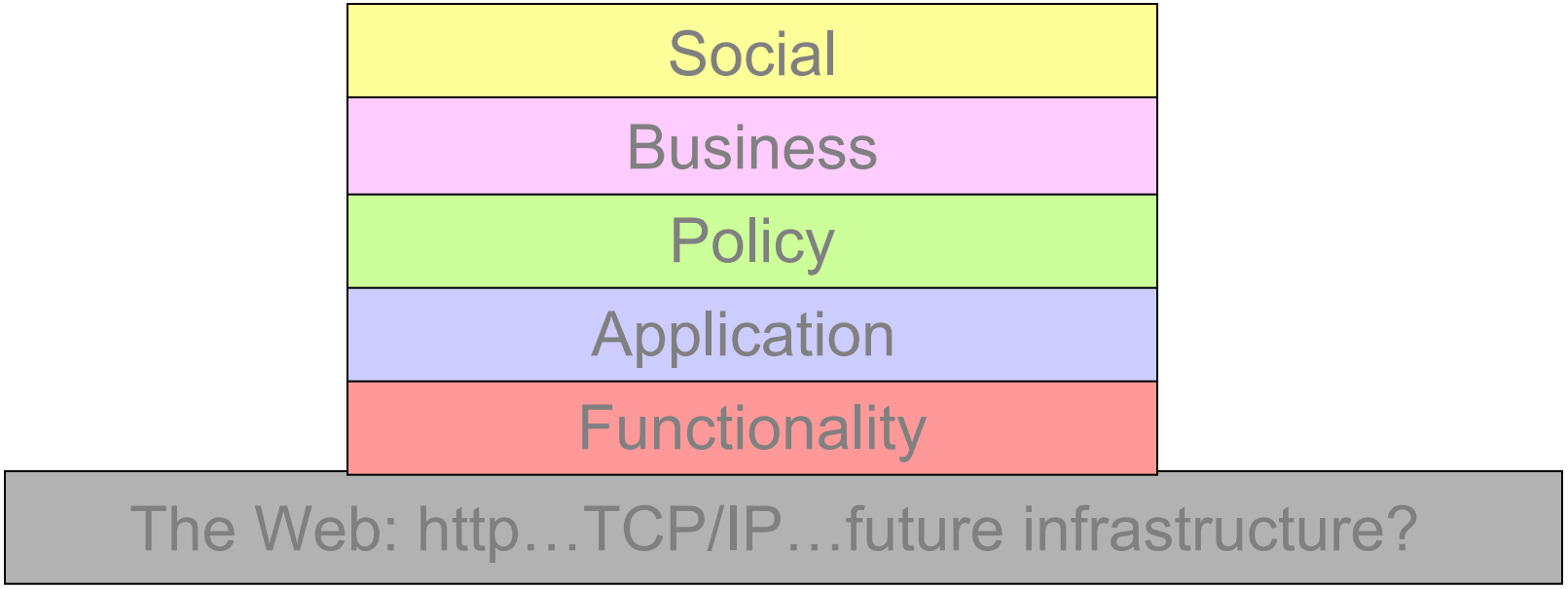
- Opaque
  - Nothing can be inferred, including sequence
  - Cannot be reverse-engineered (feature or bug?)
  - See ARCs, California Digital Library (John Kunze)
- Low-resolution date semantics
  - LCCN 99-087253
- Encoded semantics
  - ISBN 1-58080-046-7
  - Country codes… agency codes… checksums…
- Sequential Semantics
  - OCLC numbers

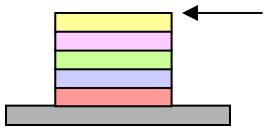
Two common pre-digital identifiers

- ISBN (International Standard Book Number)
  - Uniquely identifies every monograph (book)
  - One ISBN for each format
    - HP & SS hardback 0590353403
    - HP & SS softcover 059035342X
  - Number is semantically meaningful (components)
  - International administration (>150 countries)
- ISSN (International Standard Serial Number)
  - Uniquely identifies every serial (not issue or volume)
  - Semantically meaningless
  - International administration

The Identifier Layer Cake

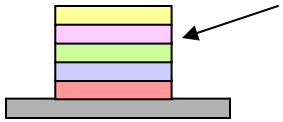
- Identifiers come in many sizes, flavors, and colors… what questions do we ask?





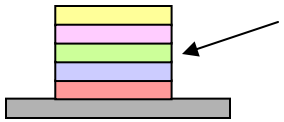
Social Layer

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



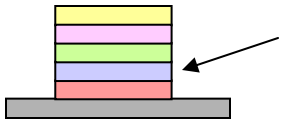
Business layer

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



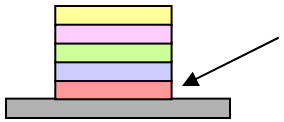
Policy Layer

- 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?
- Is there supporting metadata, and if so, is it public, private, or indeterminate?
- Is there a governance model?



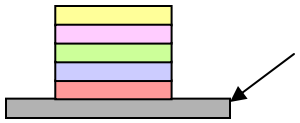
Application Layer

- 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)
- What is the means for matching persistence with the need?
- Can a given identifier be reassigned?
- Is it resolvable? To what?
- How does it 'behave'? What applications recognize it and act on it appropriately?
- Is the 'name' portion of the identifier opaque, or can it carry 'semantics'?
- Do humans need to read and transcribe them?
- Do identifiers need to be matched to the characteristics of the assets they identify?



Technology layer: The Web

Some fundamental questions:

- Must our identifiers be URIs (URLs, really)?
- 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?

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*
  - Google-like search makes non-locator *identifiers* pretty good *locators* as well

URI: Universal Resource Identifier

- Generic *syntax* for identifiers of resources
- Defined by RFC 2396
- Syntax: <scheme>://<authority><path>?<query>
  - Scheme
    - Defines semantics of remainder of URI
    - ftp, gopher, http, mailto, news, telnet
  - Authority
    - Authority governing namespace for remainder of URI
    - Typically Internet-based server
  - Path
    - Identification of data within scope of authority
  - Query
    - String of information to be interpreted by authority

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
- Partial and relative URIs

URL: Universal Resource Locator

- 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….

URL Issues

- Persistence
- 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?
- Sub-parts

Arguments for http-based identifiers

- Application Ubiquity: every Web application recognizes them. Achieving similar ubiquity for other URI schemes is very difficult
- 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

Arguments for NON http-URIs as identifiers

- Separation of IDENTITY and RESOLUTION is a small but important component of a complete naming architecture, and is poorly accommodated in current Web Architecture
- URLs make a promise: click-here-for-resolution
  - Sometimes you DON'T want resolution, or you want context-dependant action
- Not always clear what the action should be
- It is difficult to avoid branding in locators, and branding changes, threatening identifier persistance

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>

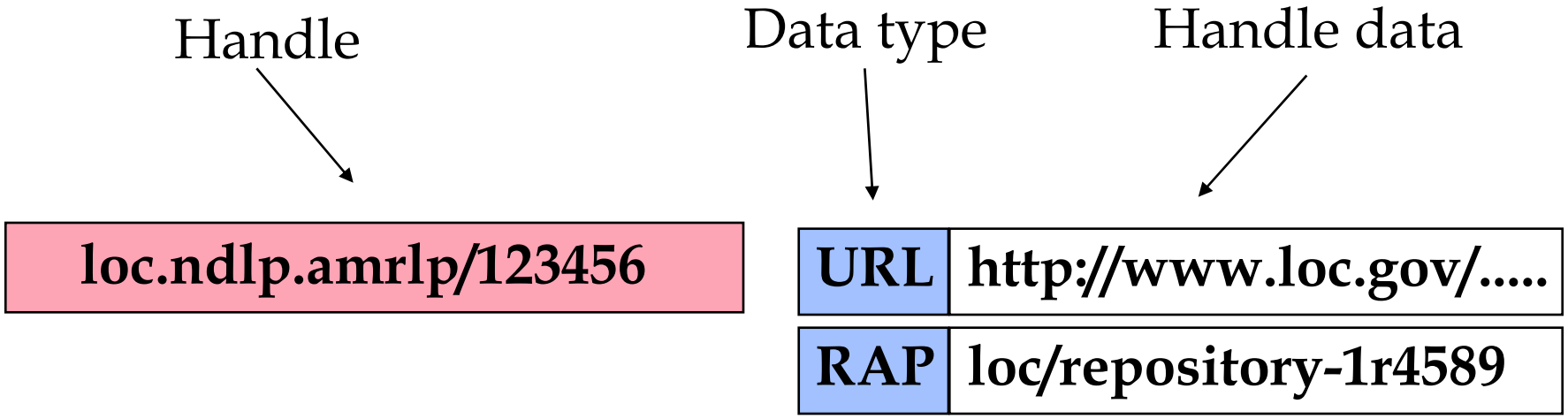
or

hdl:<naming_authority>/<locally_unique_string>

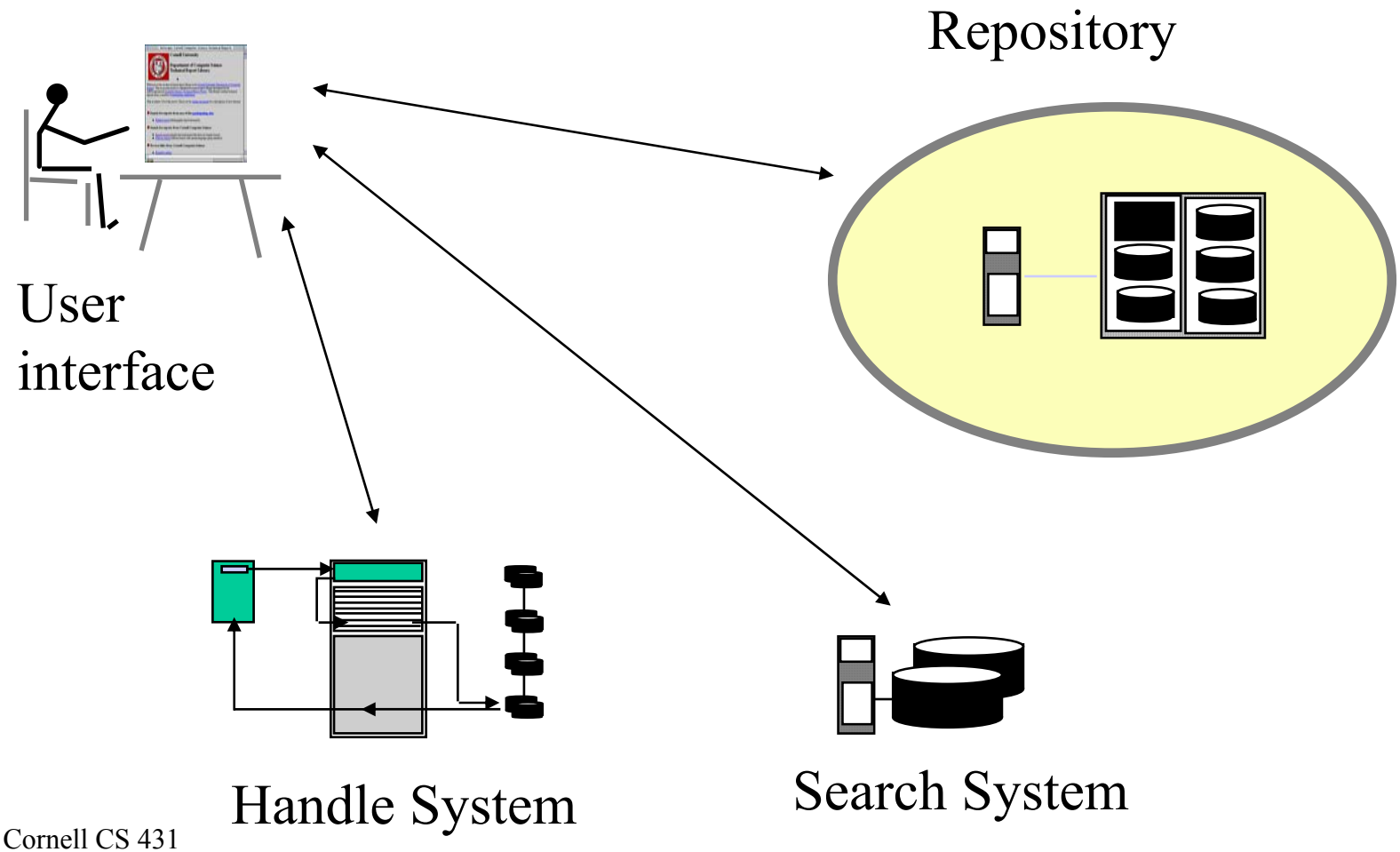
Examples

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

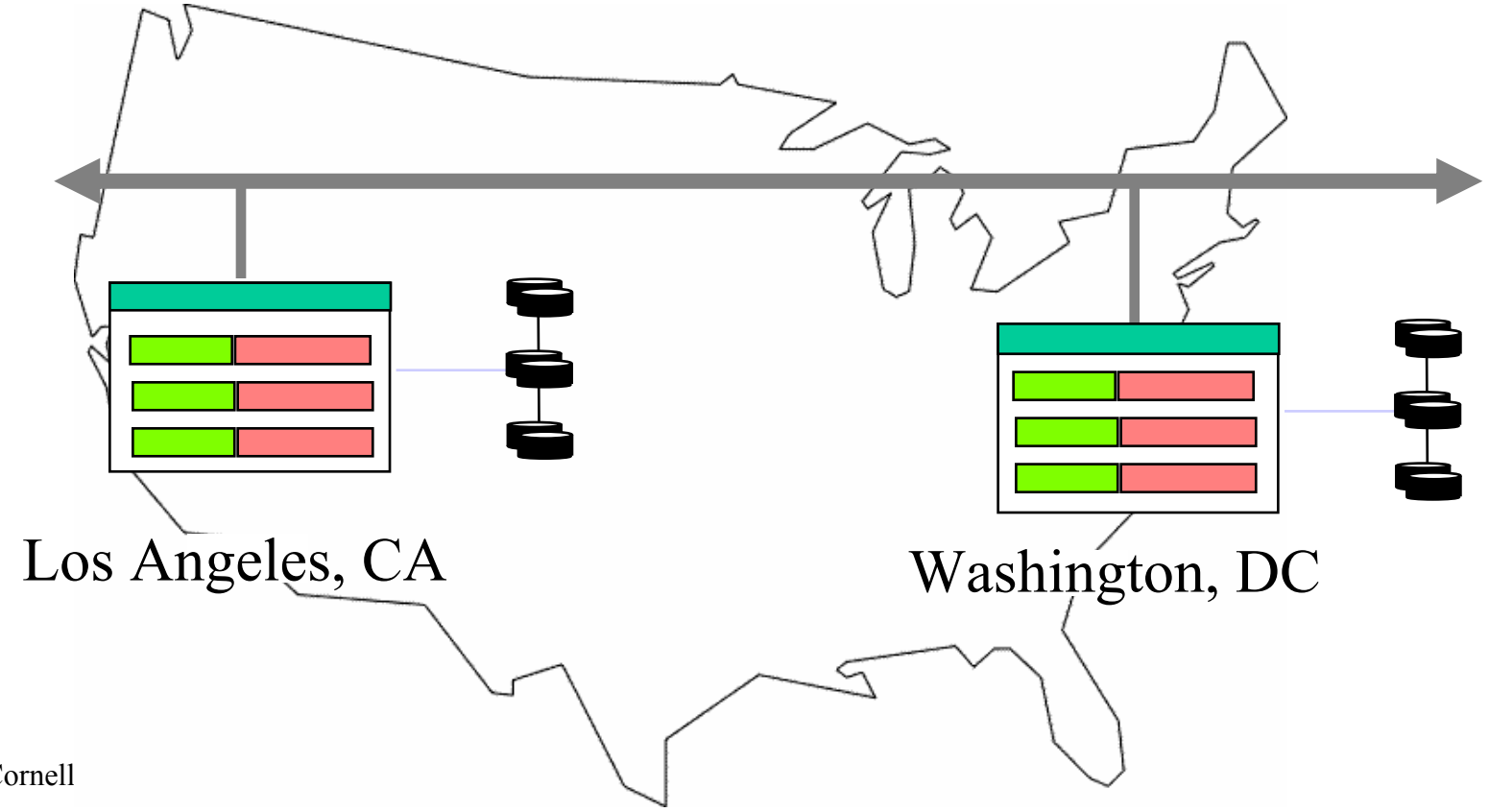


Use of Handles in a Digital Library

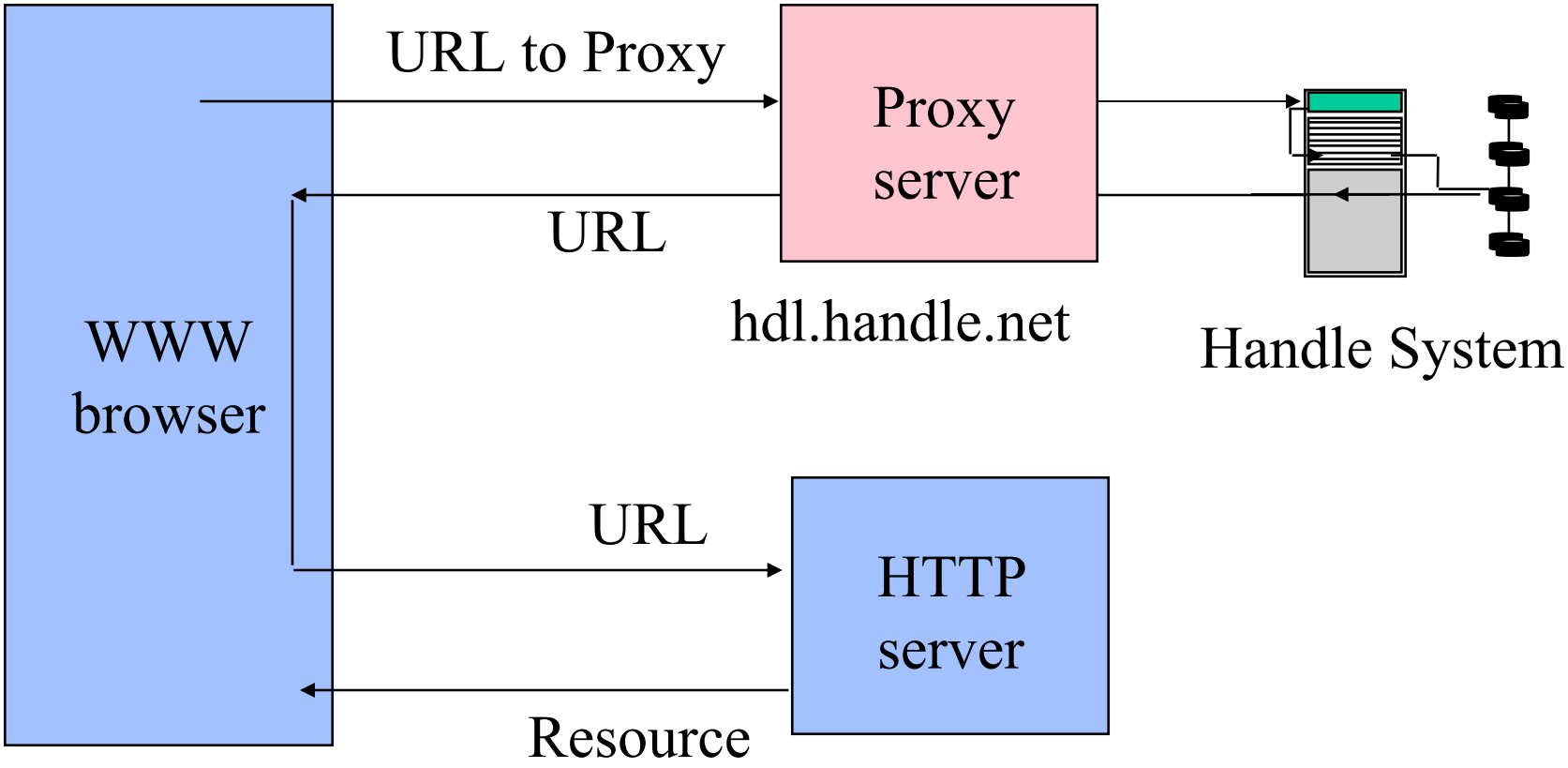


Replication for Performance and Reliability

Example: the Global Handle System



Proxy Resolution



DOI – Digital Object Identifier

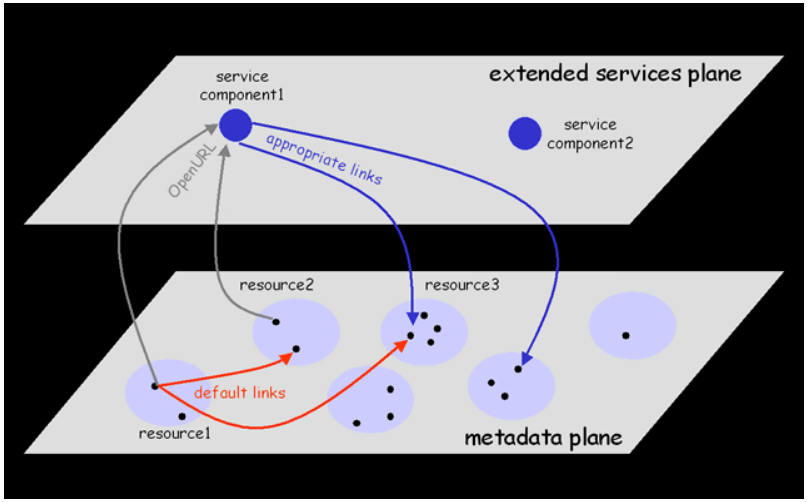
- Technology and social infrastructure for naming
- Established by publishers for persistent naming of entities (articles, journals, conference proceedings)
- Cognizant of FRBR elements
- Underlying technology is handle system
  - "persistent" names
    - Persistence is fortified by social underpinnings
    - 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?

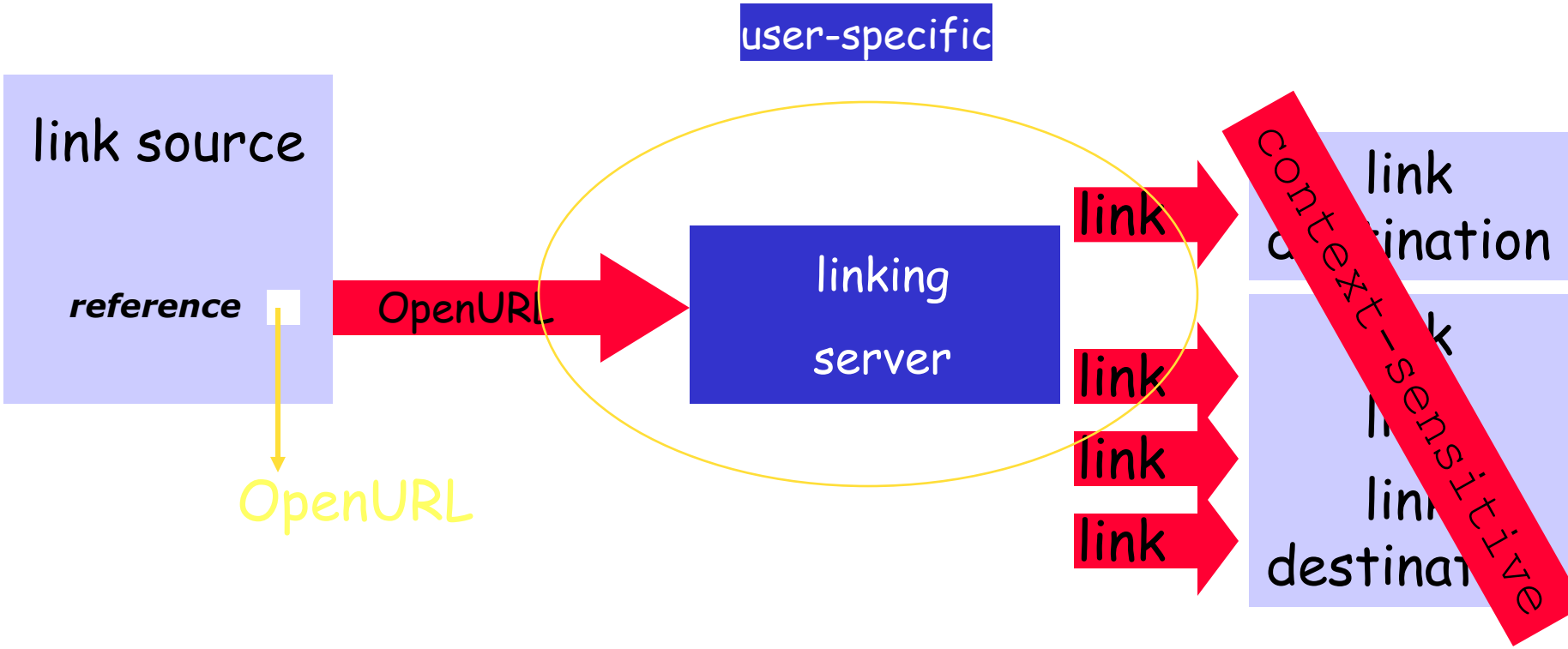
- Complexity of systems
- One size does not fit all - special purpose URN schemes have been successful, e.g., PubMed ID, Astrophysics BibCode
- No guarantee of persistence – longevity is an organizational not technical issue
- Requires well-regulated administrative systems
- Absence of "killer" applications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

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



OpenURL 0.1 syntax

http://www.mysrv.org/menu?

- id=doi:10.111/12345&
- genre=article&
- aulast=Weibel&aufirst=Stu&ISSN=35345353 &year=2001&volume=14&issue=3&spage=44&
- pid=2829393&
- sid=OCLC:Inspec

Google Scholar and OpenURL

- http://www.iwr.co.uk/information-world-review/news/2137379/google-scholar-gets-openurl-links
- "Find it at ...."

The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces

- An effort to provide a missing part of the naming architecture of the Web
- Bridge legacy identifiers and the Web
- Separate resolution from identity
- Internet Draft by Herbert Van de Sompel, Tony Hammond, Eammon Neylon, and Stuart L. Weibel
  - http://info-uri.info

What does an "info" URI look like?

- info:ddc/22/eng//004.678
  - Info: specifies the "info" namespace, or scheme
  - Namespace Token (ddc/ in this case) is a registered namespace or brand within the scheme
  - Everything that follows is at the discretion of the namespace authority that manages a given registered namespace, (and conforms to URI encoding standards)
  - No implication of resolution, though clearly services (including resolution) can be expected to emerge if "info" achieves wide use.