Issues in Managing Persistent Identifiers

Stuart Weibel
Senior Research Scientist
October, 2005
In the digital world…

- Unambiguous identification of assets in digital systems is key:
  - Physical
  - Digital
  - Conceptual

- Knowing you have what you think you have
- Comparing identity (referring to the same thing)
- Reference linking
- Managing intellectual property
What do we want from Identifiers?

- Global uniqueness
- Authority
- Reliability
- Appropriate Functionality (resolution and sometimes other services)
- Persistence – throughout the life cycle of the information object
The Identifier Layer Cake

- Identifiers come in many sizes, flavours, and colours... what questions do we ask?

The Web: http...TCP/IP...future infrastructure?
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.
• Who pays the cost?
• How, and how much?
• Who decides (see governance model)?

• The problem with identifier business models…
  • Those who accrue the value are often not the same as those who bear the costs
  • You probably can’t collect revenue for resolution
  • Identifier management generally needs to be subsidiary to other business processes
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?
• 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?
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?
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?
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

Debates about purity of *identifiers* and *locators* are ideological and unhelpful.
How we got here

- In the beginning, there was DNS
- TimBL begat URLs (within meters of where we stand)
- Uniform Resource Identifiers
  - URLs (Locators)
  - A variety of schemes, mostly grandfathered from the pre-Web Internet
  - URNs (Names, or identifiers)
  - IRIs (a URI that knows the world has more than one character set… but talk is cheap)

\[\text{URI} = \text{SCHEME, HOST, and PATH}\]

(the global file system)
<table>
<thead>
<tr>
<th>Scheme</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftp</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>http</td>
<td>Hypertext Transfer Protocol</td>
</tr>
<tr>
<td>gopher</td>
<td>The Gopher Protocol</td>
</tr>
<tr>
<td>mailto</td>
<td>Electronic mail address</td>
</tr>
<tr>
<td>news</td>
<td>USENET news</td>
</tr>
<tr>
<td>nntp</td>
<td>USENET news using NNTP access</td>
</tr>
<tr>
<td>telnet</td>
<td>Reference to interactive sessions</td>
</tr>
<tr>
<td>wais</td>
<td>Wide Area Information</td>
</tr>
<tr>
<td>prospero</td>
<td>Prospero Directory</td>
</tr>
<tr>
<td>z39.50s</td>
<td>Z39.50</td>
</tr>
<tr>
<td>z39.50r</td>
<td>Z39.50 Retrieval</td>
</tr>
<tr>
<td>cid</td>
<td>content identifier</td>
</tr>
<tr>
<td>mid</td>
<td>message identifier</td>
</tr>
<tr>
<td>vemmi</td>
<td>versatile multimedia</td>
</tr>
<tr>
<td>Interfaceservice</td>
<td>internet message access protocol</td>
</tr>
<tr>
<td>imap</td>
<td>network file system protocol</td>
</tr>
<tr>
<td>nfs</td>
<td>application configuration access</td>
</tr>
<tr>
<td>acap</td>
<td>real time streaming protocol</td>
</tr>
<tr>
<td>protocoltsip</td>
<td>Transaction Internet Protocol</td>
</tr>
<tr>
<td>tip</td>
<td>Post Office Protocol v3</td>
</tr>
<tr>
<td>pop</td>
<td>data data</td>
</tr>
<tr>
<td>data</td>
<td>dav data</td>
</tr>
<tr>
<td>dav</td>
<td>opaquelocktoken</td>
</tr>
<tr>
<td>opaquelocktoken</td>
<td>session initiation protocol</td>
</tr>
<tr>
<td>sip</td>
<td>secure session initiation protocol</td>
</tr>
<tr>
<td>sips</td>
<td>telephone</td>
</tr>
<tr>
<td>tel</td>
<td>fax fax</td>
</tr>
<tr>
<td>modem</td>
<td>Lightweight Directory Access Protocol</td>
</tr>
<tr>
<td>ldap</td>
<td>Hypertext Transfer Protocol Secure</td>
</tr>
<tr>
<td>https</td>
<td>soap.beep</td>
</tr>
<tr>
<td>soap.beeps</td>
<td>soap.beeps</td>
</tr>
<tr>
<td>soap.beeps</td>
<td>xmlrpc.beep</td>
</tr>
<tr>
<td>xmlrpc.beeps</td>
<td>xmlrpc.beeps</td>
</tr>
<tr>
<td>urn</td>
<td>go</td>
</tr>
<tr>
<td>h323</td>
<td>Internet Printing Protocol</td>
</tr>
<tr>
<td>ipp</td>
<td>Trivial File Transfer Protocol</td>
</tr>
<tr>
<td>tftp</td>
<td>Mailbox Update (MUPDATE) Protocol</td>
</tr>
<tr>
<td>mupdate</td>
<td>Presence</td>
</tr>
<tr>
<td>pres</td>
<td>Instant Messaging</td>
</tr>
<tr>
<td>im</td>
<td>Message Tracking Query Protocol</td>
</tr>
<tr>
<td>mtqp</td>
<td>iris.beep</td>
</tr>
<tr>
<td>dict</td>
<td>dictionary service protocol</td>
</tr>
<tr>
<td>snmp</td>
<td>Simple Network Management Protocol</td>
</tr>
<tr>
<td>crid</td>
<td>TV-Anytime Content Reference Identifier</td>
</tr>
<tr>
<td>tag</td>
<td>tag</td>
</tr>
<tr>
<td>Reserved URI Scheme Names:</td>
<td></td>
</tr>
<tr>
<td>afs</td>
<td>Andrew File System global file names</td>
</tr>
<tr>
<td>tn3270</td>
<td>Interactive 3270 emulation sessions</td>
</tr>
<tr>
<td>mailserver</td>
<td>Access to data available from mail servers</td>
</tr>
</tbody>
</table>
But what can you really count on?

- HTTP–based URIs (URLs) are what we can count on today

- Current URI registration procedures are unworkable
  - Scarcity of expertise
  - Techeological: strong ideologies are embedded in the process

- New URI Scheme registration standards are in the pipeline… will they help or hinder?
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
Internet Space/time continuum
Andy Powell - UKOLN

Applications that are “distant” are less likely to share understanding about identifiers. Knowledge locked within domains or lost over time or, worse, both.
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-depandan action.
- Not always clear what the action should be.
- It is difficult to avoid branding in locators, and branding changes, threatening identifier persistence.
Resolution of a conceptual asset can be problematic

- Conceptual assets should be inherently language independent:
  - Vietnamese War, 1961-1975
    DDC/22/eng//959.7043
    (English language version of DDC 22)
  - American War, 1961-1975
    DDC/22/vie//959.7043
    (Vietnamese language version of DDC 22)
Business Models may mitigate in favor of separating identity and resolution

- Content owners/managers may want to expressly decouple identity and resolution
- Appropriate Copy Problem (eg, reference linking of scholarly publishing content across subscription agencies)
- Identifiers that embed domain servers (including most http-URIs) are likely to degrade over time due to business consolidations
- URIs are global file system identifiers, and file systems change

- Web naming architectures should neither enforce nor prevent any given business model
The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces

- Internet Draft by Herbert Van de Sompel, Tony Hammond, Eammon Neylon, and Stuart L. Weibel
  - [http://info-uri.info](http://info-uri.info)
- Separate resolution from identity
- An effort to provide a missing part of the naming architecture of the Web
- Bridge legacy identifiers and the Web
- Basis for the naming architecture of Open URLs
- Is it a (registered) URI scheme?
INFO URIs (continued)

- Controversy about separating identity and resolution; IETF resistance has been substantial

- Adoption and use will determine its future – will adopters find it provides sufficient additional value to offset cost of adoption?

- Early registrants:
  
<table>
<thead>
<tr>
<th>Open URL</th>
<th>LCCN</th>
<th>DOI</th>
<th>OCLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>OCLC</td>
<td>SRW Web Services</td>
<td></td>
</tr>
<tr>
<td>Genbank</td>
<td>Fedora</td>
<td>SICI</td>
<td></td>
</tr>
<tr>
<td>Astrophysics</td>
<td>Bibcodes</td>
<td>National Library of Australia</td>
<td></td>
</tr>
</tbody>
</table>
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.
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
More Varieties

- **Domain Branding**
  - [http://elsevier.com/](http://elsevier.com/)
  - [http://pubmed.com/](http://pubmed.com/)
  - [http://LoC.gov](http://LoC.gov)

- **Functional Branding**: common behaviors established in the social or policy layers
  - [http://purl.org/](http://purl.org/)
  - DOIs
Encodings matter

- the DOI “10.1000/182” can be encoded as a URI in several ways:
  - http://dx.doi.org/10.1000/182
  - doi:10.1000/182
  - urn:doi:10.1000/182
  - Info:doi:10.10000/182
- Which of these is a registered URI?
- Which is “understood” by all Web applications?
- Which is most useful?
Recommendations and Conclusions

- Be wary (but not ideological) about semantics in identifiers
- Deviate from widely-adopted standards at your own risk (and risk to your constituents)
- There be dragons beyond the safe seas of HTTP
- Technology will not save us – Institutional Commitment is key