

OCCL Online Computer Library Center

NISO Identifier Roundtable

**Lister Hill Center
National Library of Medicine
March 13-14, 2006**

Stuart Weibel

Senior Research Scientist, OCLC Research
Visiting Scholar, University of Washington iSchool

<http://weibel-lines.typepad.com>



The Problem Statement (paraphrased)

- No shared view of identifiers systems...
 - Desired attributes of the Identifiers *per se*
 - Functional requirements for creation, management, and use (systems)
- Redundancy of effort across disciplines and sectors (many identifier systems isomorphic)
- Identifiers and identifiers systems are not the same
- Identifiers cross sector boundaries, business models, domains, and professions: interdisciplinary interoperability is important to usefulness in the Internet Commons

What we agree on:

- Global uniqueness
- Authority
- Reliability
- Appropriate Functionality (resolution and sometimes other services)
- Persistence throughout the life cycle of the information object:
 - What is the business case?
 - Persistence and permanence are not the same

Question:

Does it matter what we are identifying?

- Static documents
 - Dynamic content
 - Compound resources (multi-media, hierarchical)
 - Conceptual assets (controlled vocabularies, taxonomies....)
 - Services
 - Databases and data sets
 - Physical objects
 - Species?
 - Other?
-
- Do identifiers need to be matched to the characteristics of the assets they identify? Do Identifier systems?

Question:

Does it matter what we use the identifier for?

- Resolution (get it)
- IPR management (pay for it... or be paid for it)
- Inventory management (count it)
- Identity comparison (are two resources the same?)
- Citation (refer to it)
- Auditing (track it)
- Aggregation (bring related objects together– FRBR, compound objects....)
- Preservation (make sure it doesn't get lost)

Ideological Tarpits (1): 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.

Ideological Tarpits (2)

Opaque versus Semantic Identifiers

- Should identifiers carry semantics?
 - People like semantic identifiers
 - Semantic drift can be a problem
 - Semantics is culturally laden
- Semantic identifiers are a classic tradeoff between short term exigencies and long term problems

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://pubmed.com/...>
 - <http://LoC.gov>
- Functional Branding: common behaviors established in the social or policy layers
 - DOIs, Handles
 - <http://purl.org/...>

Ideological Tarpits (3)

Just let http do it

- Application Ubiquity: every Web application recognizes them
- 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 an important component of a complete naming architecture, 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 location-based identifiers, and branding changes, threatening identifier persistence

But what can you really count on?

- HTTP-based URIs (URLs) are what we can count on today
- Current URI registration procedures are unworkable, but are changing
 - 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?

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

Business layer issues

- Who pays the cost?
- How, and how much?
- Who decides?

- 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
 - Adding value in other ways
 - Identifier management will be subsidiary to other business processes

- An identifier system must make money, avert costs, or enable new functionality to be successful

Social Layer

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

What we need:

- A taxonomy of identifiers, identifier systems, and use cases for content types
 - Attributes of Identifiers
 - Explicit operational characteristics of identifier systems at all levels
 - Functional requirements for content types
 - Don't reinvent the wheel

But practicality and self-interest rule

- There will be more identifier systems
- Some will be motivated by new requirements, most will result from insufficient understanding of existing systems
- Political constraints cannot (should not) be dismissed
- We will have duplication, systems motivated by NIH syndrome, systems driven by self interest systems
- Systems will succeed to the extent that they solve problems, save money, create opportunity