Domain Names and Persistent Identifiers

Norman Paskin
n.paskin@tertius.ltd.uk
n.paskin@doi.org

Workshop on Domain Names and Persistence

Bristol 8 December 2011

Introduction

- Perspective: domain names in persistent identifier strings

- A disconnect exists between functionality needed for persistent identifiers and what is needed for domain names.
- But a considerable overlap: domain names are immensely useful and can meet some of the needs, but are not the optimal solution.
- Though sometimes considered as the only solution: “we have a hammer for every nail.”

- Consider a real “problem situation” rather than an abstract discussion
- URI/URN “theology” will be eschewed in favour of “pragmatic atheism” using real examples
  - Hymnsheet: IDF has a URI and URN factsheet, with contributions from W3C TAG (on URI) and IETF BoF (on URN) http://www.doi.org/factsheets/DOIIdentifierSpecs.html
- Discuss DOI system but only where relevant to domain names or to understanding identifier principles we are implementing
- We do not have all the answers! (but we do have a long experience)

Origins of DOI System

- 1994/5: internet opened up to commercial uses. Migration of primary paper-based publishing to digital networks.
- Widespread use of identifiers (ISBN, ISSN, etc.) in publishing/libraries since 1960s, proven to add value
- Domain names began to be used as identifiers – not ideal
  - 1995: *"The trouble with you publishers is that you don't know how to manage your URLs"* (W3C meeting, WKP in response to a question on this issue)
  - Publishers found this response unhelpful
- URLs do change:
  - sometimes for the wrong reasons ("uncool people")
  - sometimes for the right reasons (because they are "warm" i.e. living URLs) - more later
- 1996: 404 not found; linkrot; unstable domain names; task force formed to help publishers deal with this
  - Association of American Publishers (AAP) + International Association of Science, Technology and Medical Publishers (STM) + International Publishers Association (IPA)
- 1996: concept of DOI, creation of International DOI Foundation (1997)
  - "…unambiguous object identifier that would permit the user – or an automated process in the future – to retrieve the copyright status, owner, rights and privileges available…."
  - became DOI and developed understanding of the problem: it not digital files, it is ontology
  - selected Handle as implementation tool
  - aimed to cover multimedia, beextensible, scalable, not re-inventing standards.
  - hence "Publishers" = generic function (e.g. movies/TV assets, data, government documents…)

DOI System now

- DOI naming authorities (“Prefixes”) to date: 200,000 +
- Number of DOIs: 57 million +
- ShortDOIs: like URL shortener
- 70-90 million resolutions/month (+)
- Via Registration Agencies (independent participants in a federation) in
  - STM publishing - Crossref
  - Entertainment assets - EIDR
  - Data - Datacite coalition
  - Text publishing - mEDRA, Bowker,
  - Government documents - EC Office Publications
  - Unicode applications: China, Taiwan, (Japan)
  - (Other sectors)
- DOI®, DOI.ORG®, and doi>® are registered trademarks
- Operates as not-for-profit through shared costs
- Agnostic as to business model used by RA.

DOI System

- Not “commercial” vs “non-commercial”
  - most DOI registrants are non-commercial
- Better distinction is: DOI is for structured, large scale, management of content with the aim of aim of providing:
  - persistence of the identifier;
  - semantic interoperability;
  - some service;
  - harnessing the social infrastructure of existing organisations
- not an attempt to enforce one model, or “break the web”

DOI System

- *Persistence* = social infrastructure (aided, but not guaranteed, by technical infrastructure)
  - the most important technical component is the nut that holds the wheel

- *Semantic interoperability* : Vocabulary Mapping Framework
  - Tool based on indecs analysis; developed with funding from the Joint Information Services Committee (JISC)
  - Extensive and authoritative mapping of vocabularies from content metadata standards and proprietary schemes.
  - Not intended a replacement for any existing standards, but an aid to interoperability, whether automatic or human-mediated.
  - Includes selected controlled vocabularies and parts of vocabularies from CIDOC CRM, DCMI, DDEX, FRAD, FRBR, IDF, LOM(IEEE), MARC21, MPEG21 RDD, ONIX and RDA as well as the complete RDA-ONIX Framework.

- *Harnessing social infrastructure* :
  - e.g. CrossRef (3,600+ publishers); Datacite; EIDR; OPOCE; ISTIC

Persistent identifier principles (from ISO, indecs, DOI, etc.)

The DOI System requires that identifiers are *capable of* these functions:

- Unique identification (and description)
- Resolution
- First class naming
- Avoid intelligence in the identifier string
- Functional granularity
- Designated authority
- Appropriate access
- Metadata viewed as relationships between data
- Recognise existing schemes, allow for new schemes
- Syntactic interoperability
- Semantic interoperability
- Community interoperability
- Technology independence
- Allow any business model and services
- Subsidiarity of data management
- Do not reinvent wheels

Persistent identifier principles (indecs)

- **U**nique Identification: every entity should be uniquely identified within an identified namespace.
- **F**unctional Granularity: it should be possible to identify an entity whenever it needs to be distinguished. ["do you what you like, but say what you've done".]
- **A**ppropriate Access: everyone requires access to the metadata on which they depend, and privacy and confidentiality for their own metadata from those who are not dependent on it.
- **D**esignated Authority: the author of an item of metadata should be securely identified.

- Mnemonic: UFAD = **U**nfortunately **F**ew **A**lways **D**o

- Indecs definition of metadata: a relationship that someone claims to exist between two referents
  - stresses that such relationships can be dynamic (a spectrum of persistence)

For the purposes of this discussion, key points are: ID system must be *capable of* functional granularity and first class naming (if the registrant so wishes)

Practical Problems with domain names as identifiers

- Publishers, and their domain names, are not guaranteed to be persistent
  - e.g. Academic Press 2000
- Libraries, and their domain names, are not guaranteed to be persistent
  - e.g. National Library of Canada 2004 (changed domain name: National Archives of Canada + National Library of Canada = Library and Archives Canada (LAC.BAC))
- Some of these changes can be managed by redirects
  - others not (e.g. bankruptcy)

- Article in *Journal of Persistence* published by Acme, ID = Acme12345.
  - Journal and all back archive and records and web site sold to American Society of Persistence. IDs now have wrong embedded intelligence.
- Article in *Journal of Persistence*, ID = Persistence12345.
  - Journal changes name to *International Journal of Persisting Things*. Back archive has a different name to current stuff. (Some publishers use an abbreviation (like IJPT) in the DOI)
- multiple repositories. all with copies of *J. Persistence* articles
  - May want their own branding in identifier? ID = The Persistence Archive/ 12345.

Other issues with persistent identifiers

Also issues with using persistent identifiers, beyond domain names:

- *Assuming information that isn't intended*:
  - e.g. identifier 10.1016/1234567
  - "this DOI begins 10.1016, so I know it is published by Elsevier".
  - Wrong; all you know it was assigned at the time by Elsevier.
  - journal may have since been transferred, etc.
  - or in your university/country/region, the journal is now accessible via the Persististan National Archive....
  - ownership can be dynamic, and multiple (see Rights discussion later).

- *Imprecise definitions of referents*: "We all know what we mean by…."
  - You may, but does a third party?

- *Different services available from identifiers* e.g. Journal of Persistence article JP12345, has a data set JP12346,
  - which is assigned by another RA and has different services

Ontology issues

- Distinction of referent and resolved result
  - What is identified is not necessarily that which is obtained by resolving the identifier
  - typically a representation, an instance, class, an abstraction, a person, a thing….("Internet of Things"), not a digital object

- "Compound objects"
  - e.g. a book is simultaneously an inseparable embodiment of a work, an edition, and a format
  - which do you mean? (see "we all know what we mean by ..")
  - affects whether two such are considered "the same as…"

- May be multiple resolution destinations for an identifier
  - Resolution may be contextual

Design of domain names (by publishers etc.)

- www.New Scientist.com Domain name = main asset name (not the publisher, which is www.rbi.co.uk/)
  - lends itself to domain names/subdomains?
- www.Elsevier.com : publish around 2,000 *journals* and 20,000 *books* and *major reference works*.
  - Some have separate domain names
  - Most have sub-domain identities
  - Some may have other identities (not part of the Elsevier domain) e.g. Lancet.com (brands, historical, etc.)
  - New ones acquired/sold/deals done/transferred in or out
- Many applications now using registries which are not necessarily owned or managed by the domain name owner of the content.
- Granularity of naming can be problematic (no one right answer):
  - e.g. in CrossRef (2000), DOI naming authorities are publishers/imprints (not journals): partly influenced by early business model of DOI
- Tempting to try to brand the identifier:
  - Domain name as brand – trademark issues
    - what is brand: journal;? imprint? publisher?
  - DOI as brand
    - what is brand: DOI? CrossRef? CrossMark service?

Domain names and rights

- Identifier strings (including domain names) are the wrong way to express rights (and brands)
- Conflation of several things implied by “ownership” of an ID string (sometimes confusing identifier and referent):
  - registration of the identifier, management of the identifier, management of the referent, commercial brand, rights associated with the referent, etc.
- Separate management of the identifier string from rights
  - e.g. “Who has the current admin. rights to alter the DOI record?”
- Rights are complex, not simple:
  - Heterogeneous rights among objects under the same domain
  - One object does not have one “right”
    - *DC: Rights is at best an approximation*
  - e.g. *music:* composer, lyricist, publisher, sheet music publisher, artist, producer, recording studio; TV/film soundtrack? Pseudonym?
  - e.g. contract publishing (journals published on behalf of someone else…)
  - e.g. different territorial rights, time-limited rights, dependent rights...
  - Registrant may not wish such information to be public

Domain names as embedded intelligence

- Including a domain name calcifies a piece of intelligence into the identifier.
- At best, some information about some authority at the time of minting the identifier
  - a first place of contact for information (but increasingly fragile)
- Having explicit intelligence in an identifier string is not always a bad thing; but
- A domain name is often the wrong piece of metadata to embed in an identifier string.
  - a dynamic (and not fundamental/defining) piece of metadata
  - referent and domain name are not guaranteed to keep in step
  - one referent may be at many domains
  - it would be helpful for some uses to have the option of not embedding a domain name.
- As a principle, *first class names* are preferable: they have no embedded dependency on higher domains

Using DNS strings as identifiers:

- Brings the advantages of DNS: universal deployment, simplicity and obvious conformance with web site management structures.
- Brings the problems of domain names: wrong granularity, lack of sophistication, single hammer approach; potential for misleading assumptions re rights, registries, etc.

DOI System and domain names

- DOIs are not defined in terms of domain names: ISO Standard (ISO 26324) is an abstract specification.

- but can be optionally expressed using domain names (URN, URI)
  - compare ISO identifiers
  - e.g. *ISBN 978-0-713-99274-8* can be expressed as a printed string, bar code, QR code, URN, URI, DOI, etc.

- DOIs usually expressed as http://dx.doi.org/…
  - Sometimes explicitly, sometimes “under the hood”
  - Propagating these explicitly brings the advantages + the problems of domain names.
  - DOIs expressed as dx.doiRA.org/… would bring even more fragility
  - Propagating as DOI: is more satisfactory (but not all agree)

60 Seconds

- 23 November 2011
- Magazine issue 2840. Subscribe and save
- For similar stories, visit the 60 Seconds Topic Guide

Terminator robo-bunny

Contact lenses containing electronic displays have been placed into the eyes of rabbits to prove they are safe for use in humans. The prototype only has one pixel, but future versions could be used to display emails or video (*Journal of Micromechanics and Microengineering*, DOI: 10.1088/0960-1317/21/12/125014).

One-way only to Mars

Russia's stalled Phobos-Grunt probe has lost its chance to bring samples back from Mars's moon Phobos, Russian news agencies report. The window to fly by the Red Planet is still open, but given that the probe is not responding to commands, it looks unlikely it will make even this in time. Meanwhile, NASA is set to launch its giant Curiosity rover towards Mars on 26 November.

Monarch genes

An analysis of the genome of the monarch butterfly (*Danaus plexippus*) has identified genes important for navigating via the "sun compass" - essential aids on the insects' epic annual 4000-kilometre migration from North America to central Mexico (*Cell*, DOI: 10.1016/j.cell.2011.09.052).

Fukushima shutdown

Latest news

Comparison with trademarks?

- Domain names and trademarks – collision
  - ICANN dispute resolution allowed DNS to co-exist with existing social infrastructure of trademark principles and better able to express/use trademarks
- Domain names and identifier principles – collision
- In each case, because DNS was not devised to solve this problem/did not sufficiently consider this situation.
- Q: what would allow DNS to co-exist with existing social infrastructure of identifier principles and better enable to express/use id strings?

Conclusions?

- Can requirements for persistent identification of objects be solved whilst retaining the DNS advantages?
- Should digital network identifiers calcify in late 1990s concepts?
- Can we build by disconnecting the *Web site naming issue* from the *object naming issue* (cp. disconnect the trademark issue, disconnect the rights issue…?)
  - DOI has embraced the option of not embedding a domain name in an id string by using Handle System
- Can we fix the current system so that [some] domain name embedded has the correct functionality?

Domain Names and Persistent Identifiers

Norman Paskin
n.paskin@tertius.ltd.uk
n.paskin@doi.org