



Identifiers in the DSpace Platform

Dr. Robert Tansley
Digital Media Systems Department
HP Labs

© 2006 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

Notes

- Pertains to vanilla DSpace software and typical use as a university “Institutional Repository”
 - Users are free to customise (and some have)
- Lifecycle and other policies (e.g. whether IDs are “persistent”, persisted beyond existence of underlying object etc) are determined by institution, not DSpace software

Entities in a DSpace installation

- E-people (users, NOT authors)
- Groups (of e-people)
- Communities (e.g. research department; content type such as theses)
- Collections (Items that are somehow related)
- Items (logical works)
- Bundles (set of related files, e.g. HTML + images)
- Bitstreams (component files)
- Bitstream Formats

E-people and groups

- E-people are entities who interact with the system
 - May or may not be authors/creators
 - E-people records and identifiers independent from author/creator information in metadata
- E-people identified by e-mail address
 - Users likely to have existing, unique e-mail address
 - Both inside and outside organisation
 - Easiest way to correspond with roles databases, share between installations, etc.
 - Identified in provenance by full name and email
 - Change over time; records harder to manage, track
- Groups have no externally used ID

Communities, Collections, Items

- Generally, all given new, unique CNRI Handles on ingest
 - Some have existing Handles, which can be re-used if appropriate
 - Some have existing IDs from other schemes, e.g. DOI. Stored in descriptive metadata, can be found using search, but not resolved
- Communities and Collections are expected to change over time (items added, removed)
- Items expected to change in form over time, but not in intellectual content
- Handles not deleted when underlying object withdrawn; resolves to “tombstone”
- Items can be ‘expunged’ (completely deleted, including Handle)
- Metadata record does not get separate identifier

Handle Design Choices

- Intended to be persisted long-term despite possible changes to referent
 - Semantic free
 - Simple and free of charge (for academia), unlike DOI
 - Existing, supported infrastructure
 - Independent of format, physical location, originating organisation, transport (e.g. HTTP)
 - May be in several locations (China Digital Museum)
- Each installation has unique (sub-)prefix; generates new Handles within that
- Displayed in URL proxy form
 - E.g. <http://hdl.handle.net/1721.1/29462>
 - Feared users would not know what to do with hdl: form, and instead use URL
- Only basic use of Handle System capabilities
 - Just resolution to a URL
 - Stored, managed alongside content

Bitstreams (individual files)

- Not given Handles
- Handles are not cost-free to manage over time; could be billions of individual file Handles
- Items are expected to change over time as formats, technologies evolve (Word > PDF > PDF/A > ???); Handles identify the logical work not the set of bits that represent it
- Assumed that when people find citations/references in the future, they will be interested in the logical work
- Need to be careful what you're naming
 - E.g. does an ID for 'chapter8.pdf' refer to of bits, or the logical chapter 8?

Bitstreams (individual files)

- Each bitstream assigned unique 'sequence number' within item (1, 2, 3...)
- Currently given 'semi-persistent' URL IDs
 - <https://dspace.mit.edu/bitstream/1721.1/29462/1/ECDL+2005+final+PDF+%28Springer%29.pdf>
- URL maybe become non-resolvable; embedded Handle and 'sequence ID' can still be used to identify bitstream
- Can be used by external systems to directly retrieve the bitstream
- Proposal to move to info: URI scheme

Bitstream Formats

- Essentially identified by 'short name'
- Poor solution – varies between installations, cannot correspond between other installations and systems
- Looking to efforts like PRONOM and GDFR rather than roll-our-own
- Will still be tricky
 - E.g. AVI file with MPEG-4 video, AC3 audio encoding



i n v e n t