EZID: Easy Identifier and Metadata Management

John Kunze
University of California,
California Digital Library
USA
jak@ucop.edu

Greg Janée
University of California,
California Digital Library
USA
gjanee@ucop.edu

Joan Starr
University of California,
California Digital Library
USA
joan.starr@ucop.edu

Keywords: persistent identifiers; metadata; resolvers; API; sustainability; N2T; DOI; ARK

Abstract

EZID (pronounced easy-eye-dee at ezid.cdlib.org) is an innovative service supporting the creation and management of identifiers, their accompanying metadata, and long-term access to things on the Internet. It is one of the few services that can supply a diversity of identifier and metadata types, and do so at the earliest stages of content development, long before the content is archived or its value is understood.

EZID is run by a team within the California Digital Library (CDL), which serves the libraries of the ten campuses of the University of California, partners with national libraries, maintains the ARK identifier scheme, and belongs to global identifier organizations such as DataCite and CrossRef (FIG. 1). Started in 2010, EZID now has over 100 customers on three continents and users on all continents. In fact it is the largest and fastest growing member of the DataCite consortium. The EZID user interface is currently being revised to support multiple languages.

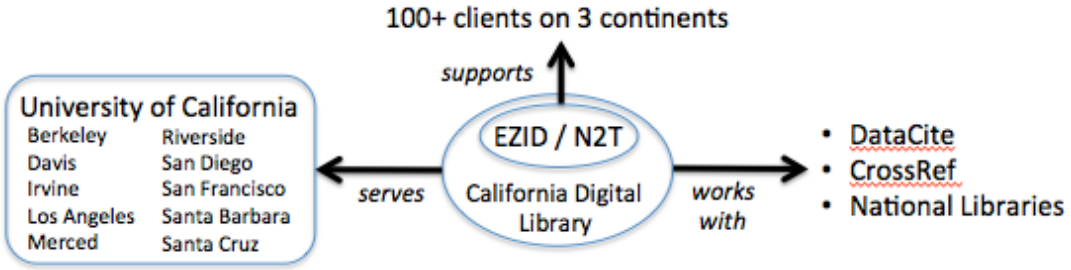

FIG. 1. Organizational context of EZID at the CDL.

EZID is unusual in supporting different kinds of identifiers. Its identifiers and metadata can describe anything of any type: documents, films, digitized maps, datasets, fossils, stars, vocabulary terms, people, etc., and it supports any identifier scheme (currently ARKs and DOIs), as well as a variety of metadata profiles, such as Dublin Core, Kernel, and DataCite.

Also unusual is that identifiers may be used for objects that are still under development. An EZID client can create "reserved" identifiers that are held privately until, for example, a draft manuscript citing them will be published. A demo mode allows anyone (no login required) to create fully functional temporary identifiers. Clients can create and use "preservation-ready" identifiers for objects that are incubating or speculative; such objects need not receive a new name when (or indeed if) they are officially published or archived, perhaps years later.

While any URL can be made persistent by carefully managing a local web server and its redirection tables, some organizations need help doing this. EZID provides them with both a user interface and an API (application programming interface) to make centralized metadata

https://doi.org/10.23106/dcmi.952137156

management easy, secure, and automatable. Every identifier has an authorized maintainer (transferrable, for example, to a successor organization) and a profile (a metadata element) guiding how all of its metadata will be presented for display, crosswalking, indexing, etc. EZID manages DOIs and ARKs that are tracked in Thomson Reuters' Data Citation Index℠.

Persistent identifiers that work with web browsers are actually URLs with carefully chosen hostnames. Sometimes a hostname identifies a "resolver", which is a special web server that forwards (redirects) public internet access requests to an object's current location, as recorded in the identifier's metadata. EZID uses two resolvers – the hostnames in these identifier examples:

http://doi.org/10.5072/FK234567 *a DOI identifier*

http://n2t.net/ark:/99999/fk456789 *an ARK identifier*

These affiliated resolvers, doi.org and n2t.net, support persistent identifier reference for any Internet user. EZID is one of the services, along with data centers and publishers, that updates DOIs at doi.org. Along with the Internet Archive, EZID also updates ARKs at n2t.net.

The N2T (Name-to-Thing) resolver at n2t.net net is non-traditional. The traditional approach to identifier persistence has been to develop a new identifier scheme and lock it up with redirection and management services designed exclusively for it. Thus the PURL, Handle, DOI, and URN schemes each has its own service "silo", and much duplicative software to manage, redirect, check links, etc. In contrast, N2T serves identifiers of any type (currently ARKs and DOIs). It is open, scalable infrastructure implemented from scratch using simple open source packages.

Traditional scheme-specific silos raise concerns for open access. With DOIs for example, it happens that any one of three specific service organizations could in theory insert advertising in or even shut down access to all mainstream scholarly journal content. EZID and N2T are deliberately scheme-agnostic, and N2T was envisioned as a resolver that could be maintained in perpetuity by a consortium of memory organizations. N2T is scalable infrastructure currently homed at the CDL and high availability is one reason CDL recently began running its infrastructure in the Amazon cloud. Until global replication across multiple regions is achieved, CDL continues to partner with EDINA to maintain an N2T replica in Edinburgh, UK.

N2T has a unique feature called "suffix passthrough" that permits one identifier for a complex object to enable resolution for many thousands of component sub-identifiers, which greatly reduces the identifier maintenance burden. Planned features in support of open linked data (semantic web) applications include "content negotiation" and a powerful inflection mechanism (short standardized extensions added to the end of an identifier).

With a view to sustainability, EZID charges a small annual fee to recover costs. Persistence is a priority, so clients that can no longer pay the fee nonetheless still retain login privileges in order to continue managing their existing identifiers. Customers include libraries, museums, archives, government agencies, publishers, and commercial data services. N2T resolver sustainability is a separate but equally important concern.

https://doi.org/10.23106/dcmi.952137156