Report on Handles, DOIs, and other Persistent Identifiers

March 2018

STIM Sub-Committee
Katie Lage | Moss Landing
Jeremy Shellhase | Humboldt
David Walker | Chancellor's Office
Andrew Weiss | Northridge

Exec summary:

This report examines the need for securing persistent identifiers for the new Samvera repository and the strategies recommended for implementing them. Among the systems and services currently available, Handles (now used in DSpace ScholarWorks), ARKs (used by CDL), and DOIs were examined. Each has their strengths and limitations. As a result, a hybrid solution may be the best approach.

This will include using Handles or ARKs as a base system for generating persistent urls in combination with a DOI-minting service (DataCite or CrossRef) for specific purposes when the need arises. Handles or ARKs would by default be assigned automatically to all repository documents, and DOIs assigned on a case-by-case basis. As DOIs are an ISO standard regularly used in online digital journal publishing, it would make sense to apply these to high-quality document types generated by the CSU – especially campus-published journals, ETDs, and grant-funder mandated open access datasets. The combined approach would provide maximum flexibility for each campus in the CSU, and would meet the needs of larger campuses as well as smaller ones.

Costs will be minimal: approximately $400 per year for Handles (system-wide) and would be free for an independent ARK (system-wide). Additionally, adding DataCite will cost as little as $3,500 for the entire CSU for up to 10,000 DOIs annually. CrossRef would be $275 plus $1 per DOI minted. On the downside, regardless of choices, there will still be considerable time and effort required to fully implement these systems in Samvera.

I. Introduction: The goal of persistent identifiers

One of the lingering problems with digital information on the Internet is the impermanence of web links. Over the lifetime of a digital object, the institution managing that object will likely change its repository software several times, each time changing the URL needed to access the object. Any existing links to the old URL will therefore break. Institutions that host open access repositories are especially in need of stable links to ensure their trusted status remains (cf. TRAC).

One solution to this problem is a persistent identifier system that lives independent of an institutional repository. Although several different persistent identifier systems exist, the basic goal and design of each is the same: The system provides a unique ID for each object and a permanent URL to access it, while also storing the most recent URL for that object. When a user accesses the permanent URL, the system looks up and redirects the user to the object's current URL. When changing repository systems, the institution need only update the current URL for each object in the persistent identifier system to ensure continued access.

In examining the landscape of persistent identifiers, we have identified the following systems as most useful: *Handles*, *Digital Object Identifiers* (*DOIs*), and *Archival Resource Keys* (*ARKs*).

II. Overview of persistent identifiers

Handles

CSU ScholarWorks currently makes use of Handles for persistent identifiers. This has been in place since 2007, when the ScholarWorks project first began. Developed by the Corporation for National Research Initiatives (CNRI) in the early 1990s, and now managed by the DONA Foundation, the Handle system consists of a central service provider, Handle.net, which provides the permanent URL for each digital object, and a local Handle service maintained by the institution, which stores the current URL for each object. When a user accesses the permanent URL for an object at handle.net, the central Handle service queries the institution's local Handle service for the current URL, in turn redirecting the user to that resources.

Subscription costs

$50 per year per handle prefix. In theory, ScholarWorks only needs a single prefix for all objects. But, in the early days of the project, separate prefixes were registered for each campus, so that the current ongoing cost is about $400 per year.

Development and maintenance costs

Hyrax does not currently have or plans to have an integration for the Handle system. To continue to mint new Handles after the migration from DSpace to Hyrax, we would therefore need to develop this integration ourselves.

There would be additional minimal ongoing effort needed to support the local Handle service. However, staying with the handle system would reduce trouble during the transition from DSpace to Samvera.

Digital Object Identifier (DOI)

The DOI system is similar to Handles in several respects -- and is, in fact, a unique implementation of the Handle system. The permanent URL for each object is available thru doi.org. Several features distinguish DOIs from Handles:

- The DOI system does not require the institution to set-up and maintain a local service. Instead, the library submits the object's current URL and descriptive metadata to a DOI registration agency -- the two most common being [CrossRef](#) and [DataCite](#) -- usually via an API.
- The institution pays both an annual subscription free and a one-time fee for every DOI it mints.
- CrossRef limits the types of objects that can be registered to (mostly) journal articles, book chapters, theses, and research data.
- DataCite is aimed at scientific and research data publishers.
- CrossRef and DataCite expose the metadata to web search engines, library discovery systems, and other online services, making objects registered with a DOI more visible to other researchers.

Subscription costs

CrossRef

CrossRef membership is $275 per year, plus $1 for each minted DOI.

DataCite

DataCite membership is $2,500 per year, plus an additional $1,200 per year per account. This includes up 10,000 DOIs per year.

The entire CSU could mint DOIs under a single account. If a campus wants their own account – in order to promote their individual identity – they would have the option to do so for an additional $1,200 per year.

Minting more than 10,000 DOIs per year would cost an additional $2,000.

Development and maintenance costs

Hyrax is in the process of developing code for minting DOIs via DataCite. We would need to develop an integration for CrossRef, were we to choose that option instead.

Archival Resource Key (ARK):

Originally developed by the California Digital Library (CDL), ARK provides the option of a more decentralized model for persistent identifiers. An institution can choose to either set-up and run its own local ARK service or register objects with the CDL's hosted EZID service.

Subscription cost

There is no subscription cost to run a local ARK service. EZID would cost $1,500 per year.

Development and maintenance costs

Hyrax does not currently have an integration for the ARK system. We would need to develop this integration ourselves.

There would be additional work to install and support a local ARK service. Currently, there are two light-weight applications available -- one in [Perl](), the other [Ruby]() -- which provide a minimal subset of ARK functionality. Neither has been updated in many years. The CDL has plans to make their ARK application code available under an open source license.

III. Analysis

As the CSU is moving away from the DSpace repository platform to Samvera, it was seen as an opportune time to evaluate the current state of the repository and its services, including persistent identifiers, since these are by no means treated the same. Analysis was initially framed by a single question: *do campuses care about persistent identifiers?* This is a valid question whose assumptions need to be examined. It is true that some campuses may be less likely to publish a large number of documents in a repository, reducing their need for persistent ids. Some will be even less likely to have a need for minting their own DOIs, given their generally more limited applications. Other campuses, at the other extreme, may have needs for these ids that extend beyond the scope of an institutional repository. As a result arguments could be made that a blanket consortium approach will satisfy no one. Fair enough. It's often impossible to achieve consensus among such a widely diverse group of stakeholders.

However, the question itself should be reconsidered in terms of the needs of a system-wide repository, not specifically the needs of the individual campuses using it. It becomes a matter, then, of *sustainable repository policy*, as outlined by the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist.* If we are to have a trusted repository, persistent identifiers are an essential component. We should as a result look beyond just local campus needs to ensure that the CSU is comparable to the standards of other institutions such as the *California Digital Library*, Harvard's *DASH*, and so on.

Moving from DSpace to the Samvera repository platform provides a few technical challenges in regard to persistent ids. The Chancellor's Office will be faced with some significant development work in Hyrax regardless of which option is chosen, as the current code for persistent identifiers appears pretty limited. To keep the existing Handles resolving, the CO would need to maintain a local Handle service and pay annual fees to Handle.net in perpetuity, even if we mint new persistent identifiers with another service. There are, additionally, some concerns about the continued viability of the Handle system and CNRI and the DONA Foundation's support for it. It is unclear if it is ultimately a good, long-term solution?

On the other hand, DOI was not really intended for some of the institutional papers and special collections we have in DSpace. It would seem somewhat unnecessary to mint DOIs for faculty senate meeting minutes, for example. It raises two questions: 1) What, if anything, should we do for these objects? 2) Does that also complicate the code in Hyrax/Samvera?

IV. Recommendation

As the group continued its investigation into the state of handles, the suggestion of a Hybrid approach seemed to make the most sense. Without a clear plan for what to do with existing handles, current campuses with large DSpace collections (and especially those with significant numbers of ETDs) would be most disrupted by the shift to Samvera. Ensuring that handles remain resolving would solve this problem. Additionally, regarding the two questions posed above, the first is relatively easy to answer, Provide easily-minted persistent urls independent of the third-party DOI minters; have a handle or ARK server available as a default approach. Limit DOIs to the types of documents they were designed to enhance, namely journal articles and book chapters; Electronic Theses and Dissertations; data sets that are mandated by grant-funders for open access.

Implementing this among campuses could be arranged as follows. First, the handles / ARKs are provided for all campuses free of charge as part of basic services provided by the CO's repository. Additional DOI minting services could be offered for campuses based on their stated need and the amount of money they are willing to contribute. Since the DOI minting services are relatively inexpensive at scale, they would provide a necessary flexibility for campuses in need of more DOIs comparted to those that have a minimal demand for them. At this point, it seems more financially viable to find a way to enter into an agreement with DataCite for at least ten campuses, which at approximately $838 per year per campus would provide more than enough ceiling for all DOI minting needs.

On a final note, it is clear that services like these in tandem with linked data services such as ORCIDs will vastly improve the quality of information infrastructure, providing stability and long-term preservation.