# Report to STIM on Handles, DOIs, and other Persistent Identifiers

Spring 2018

STIM Sub-Committee:

*Andrew Weiss | Northridge*

*Katie Lage | Moss Landing*

*Jeremy Shellhase | Humboldt*

*David Walker | Chancellor’s Office*

# Exec summary:

This report summarizes the STIM Persistent Identifier subcommittee’s examinations into the need for securing persistent IDs 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. Although each has their strengths and limitations, we see **Handles** as the best way to ensure that the content currently in DSpace will be available without disruption in the Samvera system.

It is highly recommended, though, that CSU campuses (either alone or in groups) consider adopting a DOI-minting service (*DataCite* or *CrossRef*) in addition to handles, for specific purposes when the need arises. Handles 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. This combined approach would provide maximum flexibility for each campus in the CSU, helping to meet the needs of larger campuses as well as smaller ones.

Overall costs for these systems are relatively small. It will cost the Chancellor’s Office approximately **$400** per year for Handles (an existing cost already factored in). It would be free for an independent ARK. *DataCite*, if managed centrally by the Chancellor’s Office, could cost as little as **$3,700** for the entire CSU, for up to **10,000 DOIs** total annually. An extra **$2000** would cover costs for creating more than 10,000 DOIs, resulting in costs shared at *no more than* ***$250*** *per campus per year*. Alternately, a group of several campuses managing this together without the CO could split costs for *DataCite* at a cost of no less than **$1,300-$1,800** per campus per year for up to **10,000 DOIs** *each*. Again, all libraries could share the cost burden for the whole CSU to have access to DOI minting. *CrossRef* would be **$275** **plus $.02 - $1 per DOI minted**.

On the downside, regardless of the system or systems chosen, there will be some time and effort required to integrate these systems with Samvera. However, there are already other institutions that have solved these problems that we can use as models.

# 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 ‘trust’ remains.

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](https://docs.google.com/document/d/1zNXqQNA8D4M2ZB-QOTAuv-hshJV4QcbLKH0PMCwPIPU/edit#bookmark=id.3nfd7ssyszl2) and [DataCite](https://docs.google.com/document/d/1zNXqQNA8D4M2ZB-QOTAuv-hshJV4QcbLKH0PMCwPIPU/edit#bookmark=id.q6nb94m99a) -- 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 doi content registration fees; cf. <https://www.crossref.org/fees/#content-registration-fees>. The fees vary from **($1/DOI to $0.02/DOI**) depending on type, age, and volume of DOIs minted.

#### DataCite

A DataCite membership costs approximately **$2,500** per year (**2000 euros**), plus an additional **~$1,200** per year per account **(~$600 for DOI minting; ~$600 for DOI data center fees**). This total price of $3,700 includes up 10,000 DOIs per year per account. This translates to approximately **$160** per campus per year, but might not meet all DOI needs across the whole system.

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.**

In one representative scenario, if each campus library pitched in ~**$362** per year to support five ‘official’ members, the whole CSU system could pay for a DataCite membership that provided 50,000 DOIs per year for all campuses. This is likely more than enough DOIs to meet the needs of all campuses in the CSU. See the cost chart and the details of consortium pricing in **Appendix A** below.

### 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 costs

There is no subscription cost to run a local ARK service. EZID would cost $1,500 / 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](http://search.cpan.org/dist/Noid/lib/Noid.pm), the other [Ruby](https://github.com/microservices/noid) -- 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, which exist in the DSpace system currently, 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.

# IV. Recommendation

Our exploration has raised two questions: *1) What should we do for these digital objects? 2) Does adoption of services complicate the code in Hyrax/Samvera?*

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.

Regarding the two questions posed above, the first is relatively easy to answer:

1. Keep a handle server available as a default to provide easily-minted persistent URLs independent of the third-party DOI minters;
2. Adopt DOI minter services but limit DOIs to the types of documents they were designed to enhance (i.e. journal articles/book chapters/ETDs/data sets, etc.)

The second question is a little more complicated as all of these solutions will require development by the CO and/or the repository interest groups currently working on implementation problems.

Implementing this among campuses could be arranged as follows. First, the handles 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 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 prudent to find a way in the near future to enter into an agreement with *DataCite* for at least 5-8 campuses (*see Appendix A*), which would provide more than enough ceiling for all DOI minting needs (50,000 – 80,000 DOIs per year) while being cheap enough ($300-$500 / per year / per campus) that each CSU campus could easily chip in.

On a final note, it is also abundantly clear that establishing these services in tandem with linked data systems such as ORCID, which would also be relatively cheap at $4,300 per campus, will vastly improve the quality of information infrastructure, providing stability and long-term preservation while also establishing CSU ScholarWorks as a *trusted, next-generation repository*.

# Appendix A: *DataCite* consortium costs per participating campuses



Hypothetically, if five campuses joined as a consortium, the total costs for all members would roughly be **$8,324.65** for **50,000** DOIs per year. If each of the 23 CSU campuses chipped in **$361.94** to cover the overall cost of the five ‘official’ members, the system as a whole could take advantage of this and mint up to 50K DOIs at a significant discount for all.

Obviously, we would need to determine how many DOIs would be sufficient for the whole system and plan the consortium accordingly. It seems unwise to have 23 individual members in the consortium. As we see in the table, because of the fixed DOI registration costs of **$1,189** per campus, it doesn’t make sense to have all 23 campuses as members *unless* we plan as a system to mint more than 230,000 DOIs per year.

Beyond 8 institutions, the cost per campus doesn’t decline by very much [**$1,486.55** (for 8) 🡪 **$1,292** (for 23)] As a result, consortium of 5-8 base members that allow the other 18-15 CSU campuses to mint DOIs appears to be the “sweet spot” in terms of money spent and DOIs minted across the CSU.

*Unanswered questions:*

* *What's the best size consortium to get the optimal ROI for DOI minting for the whole CSU system?*
	+ *What's the realistic number of DOIs the CSU would create as a system?*
	+ *What's the minimum # of DOIs needed for a consortium to make sense?*
* *How would this work?*
	+ *Who would manage DOI minting for all campuses?*
	+ *Central management / workflow of DOI minting: process, application, &c.*
* *How much are campuses willing to pay for DOIs?*