sciecominfo
Nordic-Baltic Forum for Scientific Communication

PERSISTENT IDENTIFIERS IN DENMARK ANNO 2012

Adrian Price

At long last the initial steps have been taken to introduce a Persistent Identifier (PID) infrastructure in Denmark. A project group, which has received support from Denmark's Electronic Research Library (DEFF)[1], has over the last year or so been working on establishing the basic building blocks of this infrastructure. The members of the project group come from the Royal Library in Copenhagen (project lead), the State and University Library in Århus, the Library of the Faculty of Life Sciences, University of Copenhagen, and the Technical Information Center, Technical University of Denmark. After considering the requirements for a PID system, which would fit well into the Danish setup and the available resources, the very basic building blocks have now been established.

I will in this article describe these "basic building blocks", as well as give a brief overview of Phase two of the project, which has also received support from DEFF and will be starting shortly in continuation of the first part of the project. There is no ambition here though to cover in any way the variety of "PID systems", standards and software available for exploiting the use of PIDs for research activity - for example DOI, Ark, Handle, the Datacite organisation, URN etc. There is ample information available, with excellent coverage of all the alternatives, so Yet Another Identifier Overview will not be attempted!

In the first stages of the project a short-list of basic requirements for a PID system was drawn up, and parallel to this the available software alternatives were investigated. In light of the available resources and other activities which were emerging in related areas of the research landscape, it was decided to use the Handle[2] software as the basis for the administration and resolving of PIDs. There were several reasons for choosing Handle: it is well-established, has a global resolving capability, has a large international user-base, and although at the moment it is managed by the Corporation for National Research Initiatives (CNRI) in the US, there are plans to find an international organisation which can take over its administration. This last point is considered important, as a PID infrastructure by nature requires a stable and permanent organisational basis. But another important factor also influenced the decision: using the Handle software would not require using a large amount of resources on software development, both in the initial establishment and in continual future development.

The actors in the PID landscape

The actors in the Danish PID landscape are for example universities, research libraries, cultural heritage and government institutions – which both publish and preserve research publications, reports and other digital objects. This digital material requires a means of ensuring permanent access as well as preservation.

Organisations and their information management systems are unstable things. When one institution gets taken over by another, are links to digital objects maintained, are even the electronic objects themselves retained? It is becoming more and more imperative that an infrastructure be established, with rules and regulations regarding the preservation of electronic material, as well as, and not least, the preservation of active links to this electronic material.

The research publications from Danish universities are a very important part of the "content" of the infrastructure. The universities are in a rather special situation: all Danish universities use the same repository software, PURE, developed by a Danish software company, Atira.[3] In a version released towards the end of 2011, it has been possible to install and configure a Handle server (in a manner similar to DSpace), which enables the issuing and resolving of Handles for local objects. Unfortunately, the details of how this "integration" between PURE and the Handle

[1] http://www.DEFF.dk/
[2] http://www.handle.net/
[3] http://www.atira.dk/en/pure/

server was implemented, was not available in phase one of this project, so exactly how the Danish universities will be brought into the infrastructure is unclear at the moment. This will be a major task to be addressed in phase two, which will begin immediately after phase one has been completed.

In the PURE-Handle implementation the administration of Handles is taken over by PURE, so there has to be established a mechanism whereby "PURE Handles" can be integrated into the central Handle resolver. Another reason that this is an important issue is the fact that this situation with parallel and independent Handle servers, will most certainly crop up again in the future: in fact it already exists, in that two Danish universities already have Handle servers which function in a similar manner. The Copenhagen Business School and Roskilde University both have DSpace, and thereby Handle servers, connected to their PURE installations.

Now that in Denmark the foundation has been laid, future phases of the project will seek to include the important actors in the landscape as a way of stabilising access and preservation in the research information environment.

Organising PIDs: danPID

As has been emphasised over and over in the persistent identifier literature, the challenges are not only technical but also administrative. Any system which relies on the implementation and maintenance of standards relies on the existence of and respect for these standards, which in turn requires an administrative setup between the partners. Another important factor of shared systems is the question of resources and financing: who will pay for maintenance and further development of services. PID systems are certainly no exception to this.

In phase two the ambition is to create an organisation which will administer the future development of the Danish Handle services. This organisation has been given the nickname "danPID". The detailed tasks of danPID are not clear at the moment, but a cooperation of member institutions, who actually contribute to content, is envisaged. The main tasks will be connected with further development of services and standards, supervision of quality requirements and keeping a watchful eye on PID developments internationally. The question of financing will be attacked from several directions, but as with all infrastructural development in the meta-research environment, resources are limited and under pressure.

The basic building blocks

In the laying of the initial building blocks, several issues have been addressed. PID terminology is an important element in communication, and is not very standardised within the PID community. Therefore a basic list containing PID and Handle terms has been produced for use in this project and will be refined as the need arises. (It is not included in this article for reasons of space, and because a number of the terms have been defined for internal reasons and are not an attempt to standardise outside the bounds of the project.)

An important part of the terminology was related to how a "PID URI" based on the Handle system should be formed. This is illustrated in Figure 1, where a persistent identifier URL and a Handle in a citation format are shown. The separate elements of this correctly formed Handle PID illustrated in Fig. 1 will be treated in what follows as they apply in this project:

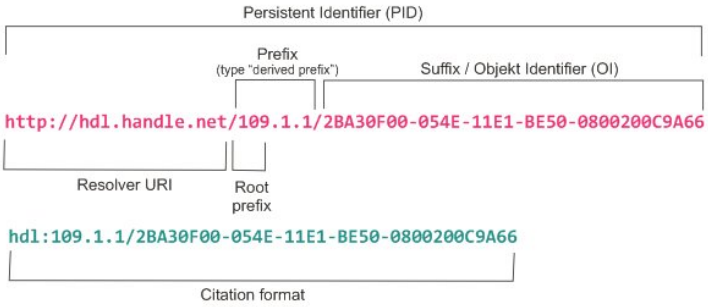

Fig. 1: Components of a Handle PID

Details regarding the individual components and rules as to how they should be formed can be seen in the Handle documentation. The relevant elements here are: the Handle prefix "109", which is the registered Handle prefix for Denmark. The project has decided that the two following positions will be used to indicate, first, the responsible institution and second, the system in question, which is owned by the institution in the previous position.

It is danPID which is responsible (or will be responsible), for issuing the first two positions ("109" and "1", the institutional position) and it is the institution which owns the system that wishes to issue

identifiers and be a part of the infrastructure, which issues the third position. This can be illustrated, as shown in Fig. 2:

Fig. 2: The derived Handle prefix

Using the example in Fig. 1 and 2:

109.1.1:
109: Denmark
1: The Royal Library, Copenhagen
1: A Dataverse system hosted by The Royal Library

Concerning the suffix, in the example used in Fig. 1 and 2, the suffix illustrated is a Universally Unique Identifier, a UUID. A UUID is a much-used type of identifier. On the initial short-list of PID requirements were a number of issues directly related to how the suffix part of the PID (the object, or unique identifier) should be formed. There are 7 conditions which apply concerning the suffix:

1. It should be possible to issue an infinite number of unique identifiers.

2. These unique identifiers must be able to be machine-generated.

3. There must be no limits as to how unique identifiers should be formed.

4. It must be possible to be able to generate unique identifiers locally (where by "locally" means the individual institutional actors in the danPID environment).

5. It must be possible to locally decide which system or method will be used to generate unique identifiers.

6. That each identifier is unique is a local responsibility.

7. It is recommended that unique identifiers are without semantics, i.e. there should not be used codes for collections, abbreviations for organisations, people's names etc. in identifiers.

A UUID has the advantage that it can be generated independently of the issuing system or organisation, which has responsibility for complying with the 7 conditions explained above. They can also be generated online or by the system where the object resides.[4] A disadvantage is that they are long (32 characters) and therefore not especially suitable for use in, for example, lists of references. Therefore it would be advisable if shorter identifiers could be created, which comply with the 7 conditions mentioned above. Systems do exist where this can be done.

As the first phase of this project come to an end, four Handle prefixes have been registered and are in operation at The Royal Library in Copenhagen and The State Library in Århus.

Conclusion

The first phase of the project is now drawing to a close, and the final report will be available from the DEFF website shortly. As already mentioned, DEFF has decided to support a phase 2, where the project group will continue with the further development of the infrastructure, expanding to include other important actors, especially universities and other public institutions.

Another important issue for phase two will be danPID: what form short danPID take, who will participate and how, as well as the vital questions regarding financing and which model will it be possible to agree on.

Adrian Price Information specialist, Library, Faculty of Life Sciences, University of Copenhagen

[4] But are there enough UUID's available? There are 340,282,366,920,938,463,463,374,607,431,768,211,456 available. " This means that 1 trillion UUIDs have to be created every nanosecond for 10 billion years to exhaust the number of UUIDs."