Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

February 2009

Relevance

CLARIN community

for all communities

Persistent Identifier Service

What is it?

A service that allows users to register and resolve persistent identifiers is the attempt to work towards stable references from electronic resources to other electronic resources or fragments. It is widely known that URLs which are currently used are not stable and will disappear within a few years. Therefore they are not suitable for references that represent valuable scientific knowledge for example. PIDs are valuable instruments to guarantee long term preservation and accessibility.

What is it for?

Users increasingly often want to

- link to a resource or resource fragment which can be a segment of a speech recording for example from a publication;
- link a lexicon entry to a part of a text in a text resource that may clarify the meaning of the lexical word;
- point from a schema of a resource to an entry in a concept registry for interoperability reasons;
- point from a metadata description to a bundle of closely related resources such as a video recording and its layers of annotations;
- carry out a semantic linking of web accessible content task by automatic methods

For all these examples it is obvious that we can foresee millions of PIDs being created and that the included links should stay for many years to support efficient scholarly work and to preserve essential knowledge.

Who can use it?

- Primarily this service was set up to serve researchers of the Max-Planck-Society and the CLARIN initiative.
- However, there are in principle no limitations to extend this service to other disciplines and to resolve huge amounts of PIDs. Special agreements would be required with the service provider (GWDG).
- The Handle System allows other institutions to also ask for a so-called Handle Prefix and to set up their own Handle Server. However, it is necessary to guarantee the persistence of such a service.

When can it be used?

The service has been set up in the form as delivered by the Handle System creators (CNRI). Mirror servers will be set up to achieve a high availability. Currently, negotiations take place with CNRI to achieve maximal independence and to implement additional functionality which is indicated as being useful within research infrastructures.

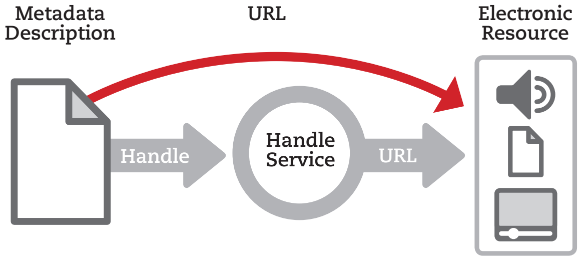
How does it work?

Persistent identifiers (PID) can be realized in different ways. The W3C Technical Advisory Group suggests the usage of true URIs which are normal hyperlinks, but which do not incorporate any physical paths or semantics that may change over time. An example is "http://www.isocat.org" since we may assume that the International Standardization Organization (ISO) will stay for a very long time and will take care that the above reference to one of its essential services will remain stable over the years even if the registration authority will change. Other communities do not trust the careful usage of URIs in the above mentioned way and therefore invented the usage of other schemes. The big national libraries agreed on using URNs to identify online-publications, however no robust and performant services are offered for open use as far as we know. The International DOI Federation created the DOI service which is based on the Handle System. It is a commercial service that allows users to register electronic resources, to associate Handles with them and to resolve them to URLs. However, for fine-grained services with millions of PIDs as they are required when working with research data the business model is not suitable and a dependency on commercial services is often not accepted.

Therefore institutions such as the Max-Planck-Society decided to set up a Handle System to offer appropriate PID services to their researchers. CLARIN will make use of this service offered by the GWDG which is one of the big computer centers of the Max-Planck-Society. Accepted repositories can request a number of PIDs by providing some information such as the URLs. In return they receive a number of unique handles which they can for example insert in their metadata records. People who are referring to a resource will include the actionable version of the PID in their resource - be it a publication or another electronic resource. Whenever a user would click on this PID a request would be directed to the PID service and the included information (URL, citation info, authenticity info, etc) would be returned, i.e. the resource could now be accessed. In the case of several copies the service could also return a number of URLs so that access optimization would be possible. Handles have a simple syntax. The prefixes are assigned to an institution with a local Handle System, the suffixes can be chosen by the local authority.

PIDs (if they are not URIs) introduce a level of indirection and complexity, since a separate service needs to be used to resolve a reference. This has the advantage that there is only one place where information needs to be changed in case that a resource is moved to another place. However, the community needs to be sure that the resolving service has an availability of 100 % and a long-term persistence.

The Handle System is used by a number of big players and it offers currently the most robust and performant PID resolution system. A number of open issues are being discussed currently with CNRI that are the creators of the Handle System. Since long term persistence of any service is difficult to guarantee it seems to be wise to include PIDs of different types (Handles and URNs for example).



Handle Syntax: $\underbrace{\textbf{1234}}_{\text{Prefix}}\textbf{/}\underbrace{\textbf{xs-45321-uvw}}_{\text{Suffix}}$

Who is responsible?

With respect to the Handle Service described in this short guide a number of institutions can be mentioned that are responsible:

- the GWDG that offers the PID registration and resolution service
- the MPI for Psycholinguistics which is responsible for the Technical Infrastructure in CLARIN
- the Handle System experts from CNRI

Whom to contact?

For the Handle System we would like to refer to the CNRI contact point: http://www.handle.net/

For all service related questions the following two addresses are relevant:

Dieter van Uytvanck (MPI): dieter.vanuytvanck@mpi.nl

Ulrich Schwardmann (GWDG): uschwar1@gwdgd.de

Where to find more information?

There is a whole bunch of information about this topic due to its enormous importance:

Handle: http://www.handle.net/

DOI: http://www.doi.org/

CLARIN-2008-2:
http://www.clarin.eu/specification-documents

IANA: http://www.iana.org/

URI: RFC 3896, http://www.ietf.org/rfc/rfc3986.txt/

URN: http://www.w3.org/2001/tag/doc/URNsAndRegistries-50

PILIN: https://www.pilin.net.au/Project_Documents/Community_Guidelines/Using_URLS_PI.htm

Responsible for the content:
Dieter van Uytvanck
MPI for Psycholinguistics
Wundtlaan 2, 6525 XD Nijmegen, NL
Website: www.clarin.eu
Email: dieter.vanuytvanck@mpi.nl

The last version of this document is maintained at the CLARIN Web-Site under documents: www.clarin.eu/documents