Common Language Resources and Technology Infrastructure

CLARIN

Persistent and unique Identifiers

2009-2-4 Version: 4

Editors: Daan Broeder, Malte Dreyer, Marc Kemps-Snijders, Andreas Witt, Marc Kupietz, Peter Wittenburg

www.clarin.eu

The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.

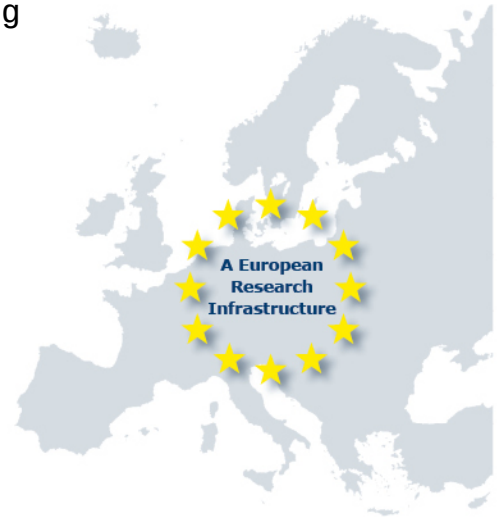
Persistent and Unique Identifiers

CLARIN-2008-2

EC FP7 project no. 212230

Deliverable: D2.2 - Deadline: 1.7.2008 (postponed to 1.10.2008 due to late start)

Responsible: Peter Wittenburg



Contributing Partners: MPI, INL, ULund, UTuebingen

Contributing Members: IDS, MPDL

© all rights reserved by MPI for Psycholinguistics on behalf of CLARIN

Scope of the Document

This document describes the goals and requirements of a registration and resolution system for persistent and unique resource identifiers that could be used by all CLARIN members and beyond, i.e. a functioning system could be used by other communities as well and there is great interest. Stepwise all CLARIN centers would need to introduce PIDs to come to a proper landscape of resources where various instances can and will be created at various places.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

CLARIN References

- CLARIN Centers Types CLARIN-2008-1 May 2008

Contents

1. Goals and Requirements for PID Systems

1.1 Goals

Within all research disciplines a continuously increasing amount of resources is produced and stored and increasingly often relations are drawn between these resources and also other resources that are essential such as references on data samples in electronic publications. For each research institution and in particular for a research infrastructure such as CLARIN it is of great relevance to develop a long-term strategy for maintaining accessibility of the resources and usability of the many references. While the relevance of the long-term accessibility of resources is without any discussion, the awareness is increasing that the created links and references are of almost equal relevance. They represent part of the knowledge that a researcher or group of researchers is building and storing, they can document important discoveries and may refer to terminology and concept registry entries which will be increasingly important for interoperability between the resources. References therefore represent substantial intellectual investments which cannot be repeated easily. Hence, for a research infrastructure there has to be a clear policy with respect to persistent and unique identifiers. For referencing to resources an appropriate granularity needs to be chosen which is dependent from the applications and discipline-specific.

In research we need to take care of two aspects: (1) We need to upload the resources into well-ordered repositories[1] which ensure of the existence, accessibility and authenticity of them to make them citable and referenceable. (2) We need to ensure that the references pointing to resources are stable, knowing that digital repositories are living organisms that are subject of continuous migration at various levels (changes in hardware and software, format changes etc). Due to all these changes we have to make sure that the identifiers used to access these resources remain valid. Although it is theoretically possible to administrate URIs[2] in a fashion that is independent of any displacement of the resource, in practice this is not done, and the literal meaning of the older term URLs[3] still applies. Often also semantics are embedded in the URL or URI, giving rise to the possible confusion in the future.

Therefore, the research domain needs other mechanisms similar to ISBN numbers for books to guarantee that references are timeless and that these numbers can be treated as the incarnation of the resources and not as one of their many copies that will exist. Only such an abstract identifier which we will call PID will be independent of place and time. Introducing such a PID, however, introduces an additional layer of complexity, since they need to be administrated and a mechanism needs to translate them into a physical location where one of the copies can be found and accessed. The definition of URIs[4] in principle opens the way to the definition of PIDs. We need to differentiate two variants[5]:

- URIs with schemes that also provide a way of locating a resource, thus coinciding with the historical definition of URLs. They point to locations, at which a web-resource can be found and accessed (compatible with an entry on a library card telling the user where on the shelves the book can be found).
- URIs with schemes that do not specify a locator service, but are intended for administrating persistent globally unique names. Historically this need was addressed by URNs, currently URNs are just one scheme of URI. The URN scheme allows the creation of many different name spaces, and every name space owner can create persistent and unique identifiers, as is done for different well known naming schemes as ISSN and ISBN. IANA (the URN namespace registration authority) keeps a list of these namespaces[6].

[1] When repositories have a long-term strategy, one can speak of digital archives.
[2] URI (Uniform Resource Identifier) used to access resources on the web. See RFC 3986
[3] URLs (Uniform Resource Locator) are specified in RFC 1738. IETF, declared the term obsolete and recommends the usage of URI (RFC 3986).
[4] IETF and W3C are busy to replace URIs by IRIs (International Resource Identifier) to adapt to the requirement of using international character sets.
[5] see also www.iana.org/assignements/uri-schemes.html
[6] http://www.iana.org/assignments/urn-namespaces/

URN type URIs therefore fulfill the basic requirements for persistency; however, they also imply the need to establish a service that translates URNs to actual URLs which is as indicated an additional layer of complexity. Such a translation (or resolution) service needs to be extremely robust, reliable and persistent since accessing a referenced resource would always require this translation step. In the following we will list a number of criteria that need to be met by such services if they want to be acceptable for the research domain.

1.2 Criteria

It is obvious that many aspects which have to do with PIDs and their resolution are of social and organizational nature. These need to be dealt with in another document. In this document we will restrict ourselves to the technical requirements.

1. Persistence and Uniqueness

At first we need to ensure that the references used are unique and persistent. While there are several suggestions for achieving uniqueness, persistency is dependent on social and organizational solutions which are not dealt with in this document, but nevertheless of great importance. It is obvious that also the resources themselves need to be persistent.

2. Contexts of References

References can occur in many different contexts (publications, web-sites, other resources etc). In each of these contexts it must be possible for the user to easily resolve them to access the referenced resources. This can be achieved, for instance, by using a special browser plug-in, that assists a www-browser to resolve a PID to its associated URI. A common practice is also to rewrite the PID into a "urlified" form that points to a "resolver service" that redirects the browser to the resource's actual location.

3. Resources and Granularity

It is the responsibility of the research disciplines and sub disciplines to find widely accepted agreements what “resources” are that should be referenceable. The requirements are highly differing, i.e. with the PID standard to be chosen no pre-definitions should be made. Important for the researchers is the possibility to not only refer to resources but also to research collections and fragments of resources[7]. The specification of fragments is very much dependent on the resource types, i.e., also in this respect the PID standard may not impose restrictions. It is the task of the repository systems or services to resolve the fragment specifications which can be for example internal identifiers. Versioning of resources is an important issue in digital repositories. It is up to them to define rules when a resource version will be given a new identifier and therefore become a new object that can be subject of references.

4. Copies

From various reasons such as load balancing and in particular long-term archiving it will be necessary to create several copies of a resource. When a user is activating a reference one of these copies needs to be accessed. Since they have the same PID their content is expected to be identical. The mechanism offering the various copies may include a ranking mechanism according to some criteria specified by the service provider.

5. Compatibility and Standards

IETF defined the syntax for URIs (Uniform Resource Identifier). Different schemas (see IANA) make use of the URI syntax specification as for example the URN schema. So if a PID string should be within a URN namespace, each PID syntax should comply with these IETF standards. Only the acceptance by these standardization organizations will improve the chance that for example web browsers will have provisions for a specific PID schema. For all non-compliant schemas the risk will be high that special plug-ins etc need to be installed to resolve references which are included in web documents. The chosen syntax should be independent of the technical implementation, since this can change within a number of years. This also includes protocols such as HTTP which may be exchanged by others at a certain moment.

[7] A mechanism can be thought of to create normal PIDs on-the-fly for selected fragments, which would bypass the need to add a fragment specification to the reference. However, such mechanism would have many disadvantages: (1) each application with help of which fragments could be identified would need to have write rights for the PID database; (2) an enormous proliferation of PIDs would be the consequence including the management and performance penalties that can be expected.

6. Additional Information

To allow users a fast interpretation of the PIDs without requiring to first look into the resource itself or explicitly invoke a metadata entry, the resolution mechanism should immediately offer limited descriptive metadata. In contrast to the PIDs these can be subject of changes. Different disciplines or institutes will associate other information in addition to the URLs and limited metadata with the PIDs. The resolution mechanism must offer these options; however, it may not be misused.

7. Semantics

It is a general agreement now that PIDs should not include semantics. Essential attributes of resources such as physical location, ownership, interpretation and grouping with others will change continuously, i.e. within shortest time semantic information in the PID would become misleading.

8. Fragment Addressing

Syntax of the PIDs as well as the resolution mechanism must accept the usage of fragment identifiers. The fragment identifiers are not part of the PIDs, however the syntax must define a delimiter and the option to add any kind of string behind the delimiter. The resolution mechanism needs to pass through the fragment specification. As far as there are widely accepted standards such as for example to specify times in video sequences these should be used to specify fragments[8].

9. Performance/Robustness/Availability

The resolution of PIDs must occur robustly and with a high performance and the services need to be available 100% of time across a very long time period to become accepted. Robustness can only be demonstrated by sufficiently long practical testing and improvement in real circumstances. High performing behavior can be achieved by a scalable architecture, by a fast network connection and choosing a sufficiently fast hardware. To prevent overhead the number of layers in a resolution mechanism needs to be limited. A high degree of availability will be achieved by providing redundancy in the architecture and by caching mechanisms. The services need to be located at institutions that have a long-term support from governments. An option for some might be using the services of the International DOI Federation that may survive for a number of years since they have contracts with publishers for example.

10. Security

Due to the importance of the data that is stored in the resolution database a high security level is necessary. Only authorized services and persons are allowed to change the database contents to protect the PID information. A regular backup needs to ensure a quick restore operation. Also storing a CRC or MD5 resource finger print with the PID, should ensure that resource authenticity can be checked.

11. Independence/Openness

The correct resolution of a PID to its associated resource is essential for the research domain and therefore a high degree of independence for a project like CLARIN is required. Models that do not allow the research domain to influence the policy will not be accepted. True independence is only given if the software that is used for the resolution is open and free of constraining licenses. Contracts need to make this clear.

12. Costs

Maintaining such a resolution mechanism and the registry facilities will cost some money. Costs emerge at both sides: the institution maintaining the resolver as well as the institutions taking care that the actual locations are associated with the PIDs. To support easy modifications APIs to the database need to be provided that can be contacted by trusted services.

In the LRT domain it is relevant to be able to refer to any resource making the number of expected PIDs very high. Therefore any business model that is linked to the number of resources and PIDs is not acceptable. Yet we don't know how the situation will develop in the various countries. Therefore, CLARIN needs to maintain its own set of PID registration facilities.

[8] Some experts don't like the combination of a PID with a fragment specification, since there are no guarantees that they will be interpreted the same way over time. But there is no alternative as has been indicated.

2. Comparison of PID Systems

2.1 Introduction

Based on the criteria presented we will compare the major PID systems that have been suggested so far. The major ones are

o	URI-URN Standard	IETF/W3C	S
o	URN Resolver	German National Library (DNB)[9]	R
o	Handle System	Corporation for National Research Initiatives Virginia	S/R
o	DOI System	International DOI Federation (based on Handle System)	S/R
o	ARK System	University of California	S/R

Here we distinguish between schemas (S) and resolution systems (R). For schemas we can only speak about a syntax specification. Resolution systems also have a software solution that transforms PIDs into real addresses.

There are a number of other schemas and resolution systems such as PURL, Info-URI and XRI. For more detailed information we refer to overviews of the MPDL [1] and the Australian PILIN [2] project.

2.2 Short Comparison Table

System	Criteria	Comments
URI-URN [4, 5]	General	Defined by a IETF Standard for the identification of web resources, yet no general resolver has been specified and developed
	Copies	-
	Standards	IETF Standard with W3C Support, the list of accepted URI Schemas can be found on the IANA Web-Site [3]
	Additional data	-
	Semantics	Left to the user/creator
	Fragments	-
	Performance/ Robustness	-
	Security	-
	Independence	All is open and freely available
	Spreading	large
	Costs	no
URN DNB [6]	General	There is a home-made resolver at the DNB that transforms standard URIs into locations
	Copies	-
	Standards	IETF compatible
	Additional data	-
	Semantics	Left to the user/creator
	Fragments	-
	Performance/ Robustness	The resolver was made for internal use only, which does not scale and not made for out of house use, not usable by CLARIN
	Security	relatively unproblematic since usage is limited to DNB
	Independence	Dependence of DNB
	Spreading	Resolver only used by DNB
	Costs	-

[9] We assume that there will be more of such home-made solutions that have a limited functionality. This is cited as one example of a URN based solution.

Handle System [7]	General	Handle system is a RFC based schema including a resolver, which has been used and improved during the last 15 years
	Copies	Supported
	Standards	Schema and protocol are specified in RFCs, yet no registration as official URI schema, for 2008 IETF acceptance as an official URI schema is intended
	Additional data	Any associated information such as metadata, rights etc is possible, database mechanism remains fast
	Semantics	Left to the user/creator
	Fragments	Plans for further implementation in 2009
	Performance/ Robustness	Obviously a software architecture that is tuned for high availability, scalability and performance, robustness has been proven by years of experience in large projects
	Security	In particular the management access has been made secure
	Independence	CNRI is open with respect to aspects of independence (mirrors, proper contractual clarifications etc) that would allow a continuation even if CNRI would stop, contracts with other institutions have been signed, the exact meaning of a patent needs to be studied
	Spreading	Not so known as URLs, but used by a number of large institutions and projects such as Library of Congress
	Costs	50 $ per year per prefix (own resolving server)
DOI [8]	General	DOI has added a business model to the Handle System and offers registration services as well
	Copies	See above
	Standards	See above
	Additional data	The INDECS schema is used for metadata, the association of other information such as rights is not intended
	Semantics	See above
	Fragments	See above
	Performance/ Robustness	See above
	Security	See above
	Independence	The DOI system belongs to a company
	Spreading	Well established in the publisher's world
	Costs	For the 500.000 objects the MPI currently has they would need to pay about 30.000 per year, since a high granularity of the references is required, costs in this size would not be acceptable. Other cost models are possible but the dependency remains and future cost control can not be assured.
ARK [9]	General	ARK comes along with an interesting schema design and a few nice features, also a resolver seems to be available, however the spreading is very limited
	Copies	Supported
	Standards	IETF draft
	Additional data	ERC (Electronic Resource Citation) metadata
	Semantics	Excluded on purpose
	Fragments	Excluded on purpose in the syntax
	Performance/ Robustness	Can't make statements
	Security	Can't make statements
	Independence	Possible
	Spreading	Little spreading as far as we know
	Costs	No

2.3 Experiences

As far as we know relatively few scientific research institutions have already experience with PID systems[10]. Some of them are Max-Planck-Institutes:

- The MPI for Meteorology has been registering larger chunks of data in the realm of their international collaboration in the climate research exchange program. The registration is done at both the DNB[11] as well as at TIB[12] Hannover, which is Registration Authority of the IDF[13]. The chunks of data are normally the chunks referred to when making a publication about climate exchange. A higher granularity of referencing is possible via the own internal PIDs stored in the internal database. A higher granularity for outside referencing would be ideal, but the current DOI model would not allow this to do due to too high costs. With the registered PIDs DOI conform metadata are associated.
- The MPI for Psycholinguistics, University of Lund and INL Leiden introduced PIDs on the basis of the Handle System in the realm of the DAM-LR Project[14]. All MPIs metadata descriptions of its about 500.000 resources got a PID entry, all three institutes maintain a local Handle Server to resolve the references and MPI mirrors the Lund PID database for testing redundancy aspects. Associated with the Handles are rights, since these should go with the objects (PIDs) and not with their instances. Until now the experiences with the Handle System were very satisfying.
- The Max Planck Digital Library needs to introduce PIDs as well, since maintaining and resolving PIDs is seen as a must for repository systems with a long-term strategy. Yet MPDL relies on URNs, but is in need of a resolution system. Together with other MPIs negotiations will be started with CNRI about fulfilling all requirements.

It is obvious that everywhere in science the registration of stable PIDs is one of the most important issues to be solved in the coming years to support stable electronic references of all sort. In the language resource domain each individual resource (even an annotation) needs to be referenced, so that we can expect a huge number of PIDs.

[10] We would like to motivate anyone to inform us about institutes with experience.
[11] This service is not available for researchers in general since the resolver was only made for internal use.
[12] Technische Informationsbibliothek Hannover (Technical Information Library Hannover)
[13] International DOI Federation – Registration of Digital Object Identifiers
[14] Distributed Access Management for Language Resources, http://www.mpi.nl/DAM-LR/

3. Recommandations

For CLARIN we will support the following recommendations:

1. All CLARIN centers will need to support persistent references to the various resources, since in an infrastructure such as CLARIN references from publications or between resources will form an essential part of the created knowledge. Therefore PIDs are an essential pillar for the language resource and technology infrastructure. In the preparatory phase we need to test this technology with a number of centers.
   Therefore, it is recommended to all (potential) CLARIN centers to make them acquainted with the requirements and solutions for creating and maintaining PIDs.
2. Essential for all research infrastructures are repositories/archives with a long-term persistency that can guarantee the accessibility of the registered resources. There is a whole variety of freely available repository systems in different states such as the US developments SRB [10] and D-Space [11] and the European developments eScidoc [12] and LAMUS [13] all with different foci. Of course, there are local database developments in many institutes that need to be analyzed in detail to check their appropriateness. It is important to note that there is no reliable PID solution without a proper repository solution.
   Therefore, we recommend to all (potential) CLARIN centers to develop a strategy to work out a persistent repository/archive solution and to get into contact with experts as soon as possible.
3. Software developers in CLARIN should consider the need of introducing PIDs. They will be included in metadata descriptions since these represent the resource objects and it will be left to the PID information to refer to the various instances of the objects.
   Therefore, we recommend taking care of the PID requirements in all CLARIN related software developments.
4. Within the CLARIN preparatory phase we should vote for one system to bundle forces. To allow centers from an early point in time to start registering their resources, we recommend to set up initial registration options at least at one powerful centre which can be used by CLARIN members. These should not be restricted with respect to granularity issues and should be ready to offer a robust and stable resolver.
   Therefore, we recommend to establish a CLARIN-wide PID registration and resolution service based on a robust system as early as possible which is open for the CLARIN community. We will check which institution can offer such a service[15].
5. Currently, we only know of one system which is performant, scalable and robust enough and that offers enough flexibility, to be used for the whole CLARIN community: the Handle System from CNRI. It's long-term existence seems to be guaranteed due to its role for the Library of Congress, for DOI and other national projects going on. However, we should achieve a high degree of independence and we should focus on standards compliance as defined by the IANA list of schemas. We should start negotiating about the necessary extensions with CNRI as soon as possible.
   Therefore, we recommend to officially starting negotiations with CNRI about the requirements which are not yet met[16].
6. The usage and meaning of PIDs in the context of the versioning is important. Although other policies are possible, it is considered important to establish as a general rule that the PID will always be associated with the original object. If a repository wishes to support a different policy, that then has to be explicitly announced in a record associated with the handle. ***Therefore we state that in the CLARIN domain a specific PID will always be associated with the original object. Deviation of this policy has to be made explicit.***
7. Of course, it is up to each institute to decide whether they want to make use of the DOI services which are also based on the Handle System. Such a decision would not harm the success of CLARIN; however, some metadata transformation will be necessary. However, we cannot recommend making CLARIN as a whole dependent of the business model of a company which is associated with costs. Due to the high granularity the costs would be too high. (Also any extra services that may be built by CLARIN on top of the basic HS, is likely incompatible with the DOI.)

[15] A German institution, GWDG, which is close to the MPG seems to be one candidate to offer such a service for CLARIN. There may be others.

[16] We had discussions with Larry Lannom, the CNRI project manager during 2008. Further talks are planned in January 2009 about the support of a Global Handle Registry mirror by the GWDG. Also support for part identifiers by the handle resolving mechanism is to be discussed.

Therefore, we recommend establishing a CLARIN PID service that is independent of any commercial business model.

8. PID services are not limited to the domain of language resources and technology, i.e. we should be open to offer such a service to others as well resp. to share such a service with others.
   Therefore, we should investigate various options of sharing a registration and resolution service with other disciplines.

4. References

[1] MPDL/FIZ eScidoc:
https://zim01.gwdg.de/repos/smc/tags/public/PubMan/Concepts/cpt_pubman_persistentidentifiers.doc
https://zim01.gwdg.de/repos/smc/tags/public/PubMan/Concepts/StoAR_PersistentIdentifiers_Version_1.0.pdf
[2] PILIN: https://www.pilin.net.au/Project_Documents/Community_Guidelines/Using_URLS_PI.htm
[3] IANA: http://www.iana.org/
[4] URI: RFC 3896, http://www.ietf.org/rfc/rfc3986.txt/
[5] URN: http://www.w3.org/2001/tag/doc/URNsAndRegistries-50
[6] DNB: http://www.d-nb.de/standardisierung/pi/pi.htm
[7] Handle: http://www.handle.net/
[8] DOI: http://www.doi.org/
[9] ARK: http://www.ietf.org/internet-drafts/draft-kunze-ark-14.txt
[10] SRB: http://www.sdsc.edu/srb/
[11] D-SPace: http://www.dspace.org/
[12] eScidoc http://www.mpdl.mpg.de/main/projects_de.htm?mp=12
[13] LAMUS http://www.lat-mpi.eu/tools/lamus