Bundesministerium
für Bildung
und Forschung

CLARIAH-DE

Persistent Identifiers in CLARIAH-DE

Summary & Best Practices (AP4.1.5)

März 2020

Version	1.0
Datum	31.03.2020
Autoren	Stefan Buddenbohm / Thomas Eckart
Prüfung:	Tibor Kalman, Antonina Werthmann
Projekt	CLARIAH-DE
Förderer	Bundesministerium für Bildung und Forschung
Förderkennzeichen	01UG1910A bis I
Laufzeit	01.03.2019 bis 31.03.2021

Table of Contents

1. Introduction[1]

This document provides a brief overview of the current landscape of Persistent Identifier solutions (PIDs) in the context of the research infrastructure projects CLARIAH-DE, CLARIN, and DARIAH. It summarizes PID systems that are actively used by project partners and the technical and organisational requirements that led to their use.

The document is not intended to be a guideline or instruction for a future unified use of Persistent Identifiers in CLARIAH-DE and its established and functioning structures. However, many years of practical PID usage in both projects allow to highlight known problems in the work with PIDs and possible strategies for their solutions. These best practices are intended to help new participants of the infrastructure. The document may also serve as orientation for further developments in CLARIAH-DE. The below described status-quo is the starting point, which has to be considered then.

2. Context of CLARIAH-DE

CLARIAH-DE[2] is the merger of the two established German research infrastructures CLARIN-D and DARIAH-DE from 2019 until 2021. CLARIN-D and DARIAH-DE are connected through their interest in the digital investigation of textual and linguistic sources from the perspective of the humanities and cultural sciences. Both networks have specialised on particular research areas for about a decade.

The aim of CLARIN-D is the establishment of a network of centres intimately entwined with particular disciplines, which shall then serve as a backbone of a research infrastructure for researchers in the humanities and social sciences in particular. The different disciplines cover a wide array of the Humanities for which linguistic resources play a major role in research.

DARIAH-DE on the other hand organizes itself as a networked community. It supports research working with digital methods and procedures in the humanities and cultural sciences with a broad research infrastructure consisting of the four pillars teaching, research, research data and technical components. Both infrastructures not only offer research-specific resources and tools but also basic services such as repositories, recommendations for standards or generic components such as virtual machines.

With the topic of basic services and generic components obtrudes the question of synergies and common standards. As important part of this undertaking, possible harmonization and/or merger scenarios for services and basic infrastructures are analysed. Mergers on an infrastructural level may not make sense in every area, but harmonization of well-established practices, as well as the use of common standards have to be considered. The user community's trust in the established infrastructures is an important asset in this regard and has to be taken into account. It is very likely that the specific approaches of the individual

[1] The authors would like to express her gratitude to Tibor Kálmán (GWDG) for his helpful comments.
[2] For more details on CLARIAH-DE (https://clariah.de/), CLARIN-D (https://www.clarin-d.net/en/) and DARIAH-DE (https://de.dariah.eu/en/) we advise to visit the according project websites.

research infrastructure reflects a certain research practice of its user community and may not be changed inconsiderately.

3. Requirements for PIDs

PI or PID[3] are both abbreviations for Persistent Identifier. PIDs fulfil a similar role for digital objects as ISBN (International Standard Book Number) for printed publications allowing the unambiguous addressing and identification of resources. As URLs (Uniform Resource Locators) are by nature not persistent and often subject to change a concept for the persistent addressing of digital objects is necessary. This becomes particularly clear with look at digital preservation or long term preservation archiving.

A concise and readable introduction into the conceptual development of PIDs, particularly the history of challenges with citing web-based references is available at *Hilse; Kothe: (2006) Implementing Persistent Identifiers – Consortium of European Research Libraries*[4].

Looking at PIDs in a more structured way, certain criteria seem to be common for the concept apart from specific implementations:

- standardisation: alignment with international standards, ensurement of interoperability among various PID solutions
- functional requirements: e.g. persistence, location independence of the PID, global uniqueness, resolvable
- flexibility, scalability: also including the possibility to add new functions without harming the system's compliance to existing standards
- compatible and technology independent
- prevalence and references on a global scale
- business model and sustainability: offerings and business strategies of the organization, e.g. how can the service be funded, by fees or other means
- policies: set of rules applied for identifiers
- organizational model: transparent decision making process and governing bodies

In the following chapter some current PID solutions are presented without claiming completeness. We also focus on PID solutions with regard to CLARIAH-DE, CLARIN[5] and DARIAH.

4. PID solutions in Use / Issuing Institutions

In the context of the CLARIN project, a survey was carried out by the CLARIN PID taskforce to get an overview of the PID solutions in use. Participating organisations of the survey use

[3] For a brief and still up to date introduction see: Neuroth, Heike/Oßwald, Achim/Scheffel, Regine/Strathmann, Stefan/Jehn, Mathias (2010): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, URL: http://nestor.sub.uni-goettingen.de/handbuch/, chapter 9.4 on Persistent Identifiers, Kathrin Schroeder.

[4] Hilse, Hans-Werner/ Kothe, Jochen (2006): Implementing Persistent Identifiers, Consortium of European Research Libraries, http://goedoc.sub.uni-goettingen.de/goescholar/handle/1/5836

[5] CLARIN and DARIAH are intentionally mentioned without their national suffixes as PIDs are not provided on a national scale but, for instance with DARIAH-DE, for the European level as a whole.

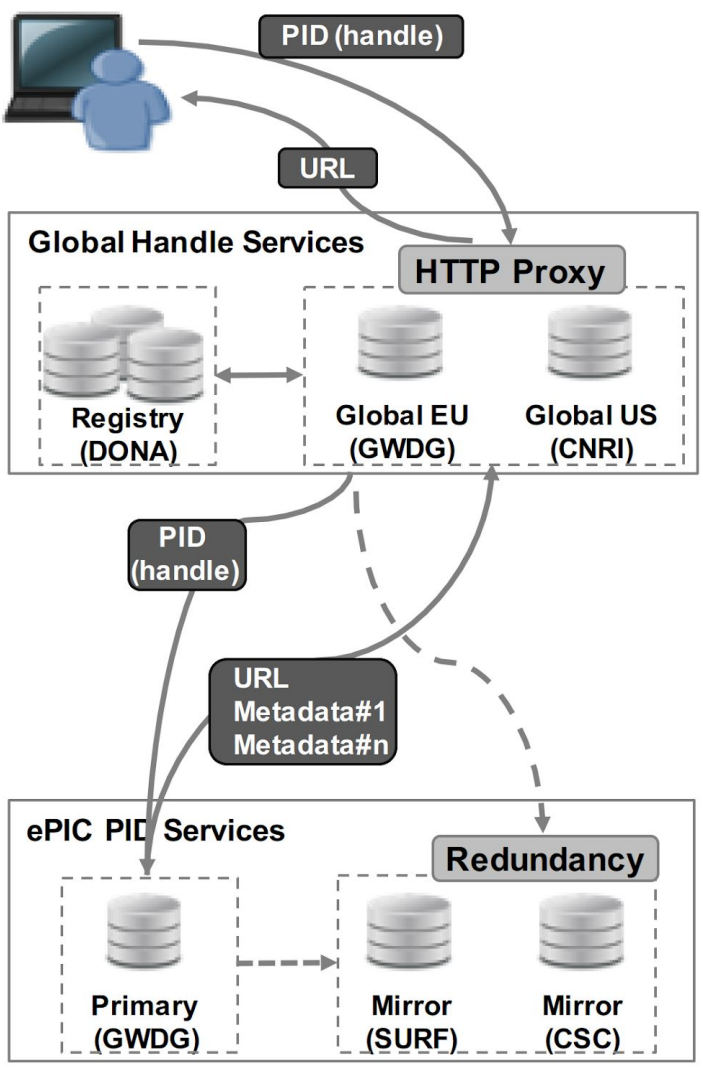
different types of PIDs to reference, identify and resolve resources. The Handle System[6] is used by all institutions that participated in the survey, but in some cases in combination with other PID systems. The following combinations were identified:

- Handle
- Handle + DOI[7]
- Handle + URN:NBN[8]
- Handle + CTS URN[9]

The most popular issuing institution for Handles is the ePIC consortium[10], which is also the chosen solution for DARIAH[11].

However, some institutions maintain their own Handle prefix and the required infrastructure. Each PID solution makes its own basic assumptions about data and their use. An advantage of ePIC PIDs is that ePIC allows the implementation of own policies, thus it prescribes comparatively few basic conditions of the referenced digital objects. By this, it is possible to assign ePIC PIDs early in the processing procedure allowing transformations of the digital object. Another advantage of the ePIC PIDs is their convertabilty to Digital Object Identifiers (DOI), as both providers use the Handle technology.

The figure presented here shows the ePIC PID implementation for DARIAH-DE. DARIAH-DE is also serving the DARIAH ERIC level, including all national branches of DARIAH. The same implementation is also used for most central applications of CLARIN.



Tibor Kálmán, GWDG, 2018

[6] https://handle.net
[7] https://www.doi.org
[8] https://nbn-resolving.org
[9] https://github.com/cite-architecture/ctsurn_spec/blob/master/md/specification.md
[10] https://www.pidconsortium.net
[11] For some details on the DARIAH PID-solution: https://de.dariah.eu/en/persistent-identifiers

Title	*Description*[12]	*Registration/Costs*	*Resolver*	*Support of free metadata fields*	*Support of fragments/part identifier/PITs etc.*
DOI	Digital Object Identifiers are coordinated since 1998 by the International DOI Foundation (IDF); URN conform service; three components: metadata, DOI as persistent identifier, technical implementation of Handle system; standardised by ANSI/NISO (Z39.84) Of importance in the DOI context is also DataCite, allowing the persistent referencing of digital objects along the research process: https://datacite.org/members.html	service is fee-based; several registration agencies e.g. CrossRef, DataCite; costs vary for registrator (DataCite: 500€ per year for up to 10,000 DOIs)	https://dx.doi.org	x	x
Handle	Handle system is a RFC based schema, which includes a resolver. Also technical basis of DOI; used by many CLARIAH-DE participants, as well as DARIAH-DE. Since 2014, the responsibility for the overall administration of the basic Handle components has been transferred to the DONA Foundation (DONA), which was constituted as a non-profit organization.	Multiple registration institutions (popular: ePIC), option to register own Handle prefix (ePIC: 45€ per year + registration; often covered by European data centres)	https://hdl.handle.net	x	x
URN:NBN	Uniform Resource Names issued by the German National Library. The partner institutions are listed here: https://nbn-resolving.org/institutions	Registration only via DNB	https://nbn-resolving.org	-	-
PURL	A persistent uniform resource locator (PURL) is a location-based URI that is used to redirect to the location of a requested web resource.	Run by Internet Archive (non profit)	https://archive.org/services/purl/ / Domain Name System (DNS)	-	o

Legend: x - 'present/available', o - 'available to some degree', - -'not supported/not available'

5. Usage Strategies

Referenced Resources

There is a variety of approaches to reference resources via PIDs. These include the decision on the reference target of a PID. Referenced resource types are:

- Metadata records
- Data files
- Collections
- (Web-)Services

[12] The table is focussing on the CLARIAH-DE relevant PID solutions. Other established solutions are not considered here but may be viewed at: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_406.pdf . The linked table also served as basis for our condensed view above.

- Publications / Publications combined with Data or Services

Creation of separate PIDs for metadata records and the referenced data files/services is actively used.

Additional features of PID services like the Handle part identifiers (fragments, templates) or storing additional metadata in the PID information record are only used by a minority of institutions.

Institutions making use of part identifiers/fragments use them for different purposes, including:

- referencing parts of a resource,
- specification of serialisation formats,
- versioning,

and more. Metadata included in the Handle system – if any – includes contact information (like Email addresses) or other resource information.

Excursus: Fragment PIDs for referencing of complex software environments[13]

An example for the citation of complex software environments has been explored within the Humanities Data Centre project in 2016. Apart from the mentioned resource types above researchers want to reference – if possible – any kind of resource. Complex software environments, like virtual research environments or visualisation frameworks, may serve as examples for these new resource types.

Although established solutions facilitating the (granular) citation of publications and (file based) research data are available, the citation of complex software environments is a not yet covered resource type in CLARIAH-DE. (Fragment) PIDs are used in this regard to reference the application or specific system state, e.g. a query term resulting in a specific data visualisation or database result. The application stands in this context as example for a type of complex research data archived in a research data centre. As said before, the usage of PIDs is common to ensure stable references to publications and is increasingly accepted for file-based data too.

As the ePIC PID[14] service offers the creation of Fragment PIDs, it may be used to reference certain application or system states. Citing from the article of Bingert/Buddenbohm: "These (Fragment PIDs) are identifiers that not only can be resolved to a given location, but also allow to forward parameters when the PID is resolved. With this, it becomes possible to make a stable reference to defined system conditions – but it also requires some effort on the side of the referenced system. The Fragment PID can be used to present the user predefined configurations of a system. This could for instance be a visualisation with specific parameters or a simulation of a specific state. As mentioned in the introduction PIDs are a common means to cite a very broad range of various objects. In the humanities PIDs are used to identify collections, content or objects. PIDs are not only able to reference to

[13] Bingert/Buddenbohm (2016): Referencing of complex software environments as representations of research data. DOI: https://doi.org/10.3233/978-1-61499-649-1-58

[14] FAQs regarding ePIC PIDs: https://www.pidconsortium.net/?page_id=1060

definite objects but may also reference object fragments with the usage of a Fragment PID. This may be passages of text or illustrations or links to certain sections in digital media by following examples:

http://www.domain.org/book1@page=10

http://www.domain.org/video1@begin=10&end=20

where in this example book1 and video1 represent the PID. The naming schema for such PIDs differs between the existing PID systems and is not subject of discussion in this paper. But important to note is that with those identifiers an unlimited number of fragments in an entity can be referenced and provide the level of granularity that is necessary for scientific citations."[15]

Current citation practices in the Social Sciences and Humanities

Becoming crucial not only for the SSH in the near future is the ability to link datasets and more traditional publications such as papers. This can be described as dynamic data citation, when one also takes into account data coming from sources such as Twitter. A very concise and up to date summary of current citation practices in the social sciences and humanities is available from the SSHOC[16] project. The inventory of citation practices came, apart from a landscape overview of current standards and best practices to the following conclusions[17]:

- the use of explicit versioning policies requiring an explicit specification of the data object version or the lack thereof;
- the use of a "tombstone" landing page associated with PIDs for deleted data object
- support for citing parts of complex objects i.e. using PIDs that support part-identifiers or some other suitable technology, that is able to efficiently access parts of larger objects;
- some citation metadata seem to be overly extensive, though we cannot judge the absolute necessity of such a practice. However, we do want to recommend using DataCite citation metadata as a minimal set and keeping the citation metadata manageable and with a predictable structure;
- increase the actionability of citations and landing pages: using actionable identifiers and well - structured and informative landing pages. This is also consistent with the DataCite landing page recommendations;
- it would be useful to implement interoperability services in the publication platform: for instance, providing standard harvesting protocols that could improve the possibility of sharing metadata with other platforms, providing a well-documented API to access data can simplify machine-actionability for identifiers, etc.

[15] Bingert/Buddenbohm (2016), p. 63
[16] https://www.sshopencloud.eu/
[17] The listed conclusions cite directly Larrousse/ Broeder et al. (2020), p. 27

Resource Updates

Another diverse field is the handling of resource updates, like the deletion of resources or when a new version of an existing resource is published.

Many institutions do not allow the deletion of published resources. In cases, where resources are deleted from a repository, popular strategies include:

- Keep object PID
- Keep PID + keep metadata
- Keep PID + adapt metadata

In cases of resource updates, different strategies are in use of which the most popular include the systematic creation of a new PID for the new object, often combined with the creation of a new PID for the new metadata record.

6. Discussion

This CLARIAH-DE report on PID best practices offered a brief overview of the current practices and standards from a SSH angle, particularly taking into account the situation within CLARIN-D and DARIAH-DE.

Currently (2020), both infrastructures use ePIC PIDs. CLARIN-D is a direct PID consortium member, whereas DARIAH-DE is represented by the GWDG, a long standing infrastructure partner for DARIAH. Of course, it is possible for researchers and projects in the CLARIAH-DE universe to use other solutions for their research outputs. ePIC PIDs are the solution of choice, when it comes to added functions of services like the DARIAH-DE repository. An ePIC PID is assigned to each item in the repository. On the other hand, the non-mandatory recommendation of ePIC PIDs may pave the way for a heterogenous PID landscape within CLARIAH-DE. An established example for this aspect may be seen in DataCite. DataCite is offering its consortium members DOIs and, for instance, a partner in CLARIAH-DE may also be member of this consortium[18] beside the ePIC PID consortium membership mentioned above.

As long as the use of PIDs complies to the established standards, the above mentioned situation may not be a problem at all. Certification initiatives like the Core Trust Seal or the DINI Certificate might serve as orientation in this regard. Both offer reputable quality seals for research data repositories (CTS) or publication repositories (DINI). Beside this, internal assessment or compliance checks are available, such as the CLARIN Centre Assessment[19], a mandatory instance to pass for becoming a CLARIN centre[20].

7. References

[18] This is the case for the Göttingen State and University Library, which is a DataCite partner but - over the bypass of the GWDG - is also an ePIC partner.
[19] https://office.clarin.eu/v/CE-2013-0095-B_checklist-v7_3_1.pdf (section 7)
[20] https://www.clarin.eu/content/overview-clarin-centres

- Bingert, Sven/Buddenbohm, Stefan (2016): Referencing of complex software environments as representations of research data. IOS Press. DOI: https://doi.org/10.3233/978-1-61499-649-1-58.
- CLARIN Report on "Persistent and Unique Identifiers" (2008), URL: https://www.clarin.eu/file/1385.
- CLARIN Usage survey among all CLARIN B-centres carried out by the CLARIN PID taskforce during Q3/2019 (unpublished).
- Hilse, Hans-Werner/Kothe, Jochen (2006): Implementing Persistent Identifiers - Consortium of European Research Libraries, PURL: http://resolver.sub.uni-goettingen.de/purl?gs-1/5836
- Kálmán, Tibor (2018): Use of ePIC PID-Services at the GWDG. PID: https://hdl.handle.net/21.11101/0000-0007-E17E-E
- Larrousse, Nicolas/Broeder, Daan/Brase, Jan/Concordia, Cesare/Kalaitzi, Vasso. (2019). SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning (Version v1.0). Zenodo. DOI: https://zenodo.org/record/3595965
- Neuroth, Heike/Oßwald, Achim/Scheffel, Regine/Strathmann, Stefan/Huth, Karsten (Ed.)(2010): NESTOR-Handbuch zur Langzeitarchivierung, Version 2.3 (2010), URL: http://nestor.sub.uni-goettingen.de/handbuch/