DRAFT VERSION – only intended to be used as background for the CLARIN PID workshop

Data Service Infrastructure for the Social Sciences and Humanities

ECFP7
Grant Agreement Number: 283646

Deliverable Report

Deliverable: D5.1a
Deliverable Name: PID Services Report
Deadline: M24
Nature: R

Responsible: Birger Jerlehag, UGOT
Work Package Leader: Daan Broeder, MPG-TLA

Contributing Partners and Editors: Timo Gnadt, UGOE; Arjan Hogenar, DANS; Bart Jongejan, UCPH; Merja Karjalainen, UGOT; Przemyslaw Lenkiewicz, MPG-TLA; Jens Ludwig, UGOE; Catharina Wasner, GESIS

Contents

1. Executive Summary

In DASISH (www.dasish.eu) T5.2 the aim is to promote the usage of Persistent Identifier (PID) services at data centres within the DASISH communities (CESSDA, CLARIN, DARIAH, ESS, and SHARE), all within the Social Sciences and Humanities. The reason to encourage the use of globally unique persistent object identifiers is that too often individuals are confronted with objects[1] that are no longer traceable and accessible on the Internet. In most cases the reason for the disappearance of objects is the absence of a policy for sustainable access within the organisation responsible for the production of these objects. Without such a policy, which should include the use of PIDs for unique and sustainable identification, objects may be deleted or re-located without alerting users.

The necessity of using globally unique Persistent Identifiers (PIDs) should therefore be obvious to researchers and research organisations as well as to people responsible for data archives and repositories. These identifiers are crucial for the advancement of science, as they are (or should be) coupled to policies on sustainable access.[2, 3] The need for PIDs has driven the development of PID systems, for example The Handle™ System.

There is a need for additional services coupled/related to the registration of PIDs for objects, for example services for PID registration and possibilities to store descriptions of the objects in central PID metadata repositories. However, user requirements for PID service providers may differ within scientific or scholarly disciplines. Consequently, it is important to assess these requirements within the different communities to find out if it is possible to arrive at a general, widely accepted list of requirements. To find out what these requirements are a survey was conducted among the data centres and communities of the SSH infrastructures composing DASISH (Questionnaires in appendix 1 & 2).

This report focuses on the answers from these surveys, and includes a more thorough description and comparison of three of the most commonly used PID services within the communities. The descriptions of the PID services have been verified by the service providers to ensure their correctness. The PID service providers are analysed and compared in the light of the requirements derived from the surveys of the DASISH communities (see Conclusions chapter 3.3). One desirable outcome of DASISH T5.2 would be if these PID service providers

---

[1] "A Digital Obje[c]t is any kind of digital resource, which is identified by at least one PI assigned by a trusted PID." APARSEN-REP-D22_1-01-1_9 (2012), p 36 http://www.google.com/url?q=http%3A%2F%2Fwww.alliancepermanentaccess.org%2Fwp-content%2Fplugins%2Fdownloadmonitor%2Fdownload.php%3Fid%3DD22.1%2BPersistent%2BIdentifiers%2BInteroperability%2BFramework&sa=D&sntz=1&usg=AFQjCNG9cMI9K6w6FsrSAEk3_tpsomp7GQ

[2] The Decay and Failures of Web References (2003). Draft version: http://www.dmst.aueb.gr/dds/pubs/jrnl/2003-CACM-URLcite/html/urlcite.html. Published version: Communication of the ACM, 46 (1):71-77. doi:10.1145/602421.602422.
[2]

[3] Towards Persistent Identification of Resources in Personal Information Management (2013). http://ceur-ws.org/Vol-1091/paper7.pdf. Proceedings of the 3rd International Workshop on Semantic Digital Archives (SDA 2013), p 73-80.

would take action to further improve their services based on this analysis and the recommendations derived from it.

The report also discusses PID service functionality above the simple DO URI resolving. Recent developments[4] in thinking about data management for research data have also indicated a need for some tightly coupled metadata linked directly with the object identifier. For instance, for integrity checking of data objects by using checksums. PID services directly supporting such tightly coupled metadata can be considered advantageous.

The report will also serve as a status snapshot of the major players in current EU PID service landscape, indicating the availability and trustworthiness of the respective services so it can be used to make a choice for centers needing such service.

The conclusions drawn from this task will be communicated to the DASISH communities, and recommendations and guidelines will be disseminated to the data centers by way of collaboration with DASISH WP7 – "Education and training". They will also be communicated to the service providers to inform them about the DASISH requirements.

[4] See for instance the RDA DFT WG and PIT WG discussions.

2. PID Services

DASISH T5.2 focuses on currently operational PID services, i.e. services provided by external organisations that at the very least offer Identifier to URI resolution and require only minimal actions and no provisioning of any own services of the object hosting sites. The external service providers should be arguable stable and persistent, of course. Any extra functionality, such as the possibility to provide with the PID tightly coupled metadata, for instance for use in an integrity checksum, will be explicitly mentioned.

There are a number of service providers using possibly different PID systems (see the CERL report in Appendix 3). The focus within DASISH T5.2 is on three PID service providers that already play a central role within the five DASISH communities: DataCite, the European Persistent Identifier Consortium (EPIC) and the URN:NBN Cluster PID service hosted by the German National Library. These services are described in the following subsections. Any impediments for general use, such as limited coverage, will be explicitly mentioned. Evaluations with respect to performance aspects such resolving speed, and generation of possibly very many (>millions) of PIDs is beyond the scope of this report. But information about such aspects does appear when it mentioned in user-surveys.

The offerings of the services have not been tested in practice in a methodic way. They are however used by the Home institutions of members of the Task group, so there are practical experiences of them. The descriptions of the services are built on information found at the service providers' homepages and other network resources. To ensure the correctness of the descriptions they have been commented upon, updated and verified in Quarter 4/2013 by representatives from the different service providers' organization.[5]
.

2.1. Towards a European DOI-Based PID Service – DataCite

Introduction

DataCite is a not-for-profit international organisation formed in 2009[1]. The organisation consists of a managing agent (currently German National Library of Science and Technology[2]) and Members[3], and is represented by a board[4].

DataCite operates globally, but is nationally represented to reach out to research groups in different countries. Member institutions interact directly with clients[5], and offer services to the research community. Member institutions that provide identifiers for clients are called *Allocation*

[5] DataCite: Brigitte Hausstein (GESIS); EPIC: Ulrich Schwardman (GWDG); URN:NBN: Nicole von der Hude (DNB)

Agencies. In most cases, there is one Allocation Agency per country. *Associated Members* are not Allocation Agencies.

Mission Statement

DataCite's mission is to [1]:

- establish easier access to research data on the Internet
- increase acceptance of research data as legitimate, citable contributions to the scholarly record
- support data archiving that will permit results to be verified and re-purposed for future study.

DataCite's Business Model Principles states that "DataCite is an international association dedicated to making it easier for everyone to identify, cite, discover, and use research data" [2, p. 1].

In fulfilling its mission, DataCite focuses on working with data centres and organisations that hold data.

Organisation of DOI PIDs

DataCite uses DOIs (Digital Object Identifiers) as persistent identifiers. In the DOI handbook[6] a DOI is defined as a digital identifier of an object, rather than an identifier of a digital object. The DOI system is managed by the International DOI Foundation (IDF), and is an extension of the Handle System architecture (see more about IDF's role in the section about Governance Structure). Technologically speaking, DOIs are just handles and the resolving mechanism is provided by the Handle System infrastructure just as in the case as EPIC.

When registering a DOI for an object, a DOI name (prefix/suffix – e.g. `10.1000/182`) is assigned together with location information (such as a URL) for the object. Additional metadata that describe the object in more detail is submitted to a metadata repository. The DOI name concatenated with `http://dx.doi.org/` forms an actionable URL, e.g. `http://dx.doi.org/10.1000/182`. When clicking on such a URL, one is redirected by a resolver[7] to the specified location of the object. The complete URL can be embedded in a document, in which case the DOI is indicated as follows in the citation: `doi:10.1000/182` which string, dependent on the document viewer can be an actionable string. Otherwise the PID can be embedded in a URI (URLified), such that it becomes actionable in most document viewers. If an object with a registered DOI changes its location, the location information has to be changed in the register to maintain the accuracy.

The global uniqueness of the centralized allocation of DOI names is secured by using prefixes to create unique namespaces that can be used within different organisations. Suffixes are assigned by the registrant/organisation, and are unique for each object within that organisation. Prefixes are distributed by Allocation Agencies that are Members of Registration Agencies, such as DataCite (see the section about Governance Structure).

The additional metadata, one of the features that distinguish DOIs from “pure” Handles, are stored at the Registration Agencies (e.g. DataCite Metadata Store). The metadata are used for e.g. the service provided by DataCite, through which searches for related data can be done.

Governance Structure

The International DOI Foundation is the registration authority for the ISO standard (ISO 26324) for the DOI system. IDF is also the governance and management body for a federation of Registration Agencies (RAs).

There are several RAs for registering DOIs[8], since different communities require different services. They all provide services that allocate DOI name prefixes, register and resolve PIDs, and store metadata about the objects. DataCite is one of the RAs in IDF. DataCite Members, in turn, can be Allocating Agencies that allocate DOI names on behalf of the DOI Registration Agency of DataCite.

Services Offered by DataCite

Some examples of services offered by DataCite[9]:

- **DataCite Metadata Store** is a service for data publishers to mint DOIs and register associated metadata.
- **DataCite Metadata Search** is a search service based on metadata for datasets registered with DataCite.
- **DataCite OAI Provider** exposes DataCite Metadata for harvesting (OAI-PMH).
- **Test Environment** is set up for testing DataCite’s services, including the DOI registration. The test environment is a closed system.
- **Content Negotiation –** DataCite's Content Resolver exposes the metadata stored in the DataCite Metadata Store (MDS) using multiple formats. It can also redirect to content hosted by DataCite participating data centres. It is therefore possible to access data directly by using a DOI. Furthermore, DataCite joined forces with CrossRef to establish a working HTTP Content Negotiation.
- **DOI Citation Formatter** – set up in collaboration with CrossRef, and creates different citation formats for DataCite and Crossref DOIs. Users can choose from more than 500 different citation formats in 45 different languages.
- **DOI Statistics –** provides statistics of DOI registrations and DOI resolutions, filtered by Allocator, Datacenter, or Prefix.

Requirements – DataCite Client Responsibilities

When signing a contract with a DataCite Allocation Agency, the clients have responsibilities according to the DOIs and the objects that are assigned DOIs [2]. The main responsibility is that the clients **commit to data persistence**. This means that the clients are expected to store and manage objects so that persistent access is provided. Maintaining all URLs that are associated with the DOI is included to the data persistence commitment.

For clients that are registering DOIs for data, there are some requirements regarding metadata and landing pages for the DOIs, for example (for more detailed information, see [2]):

- **Metadata** – the client has to provide at least mandatory metadata, and share their metadata for use in various DataCite services, e.g. for discovery purposes.
- **Landing Pages** – the landing page is the web page that the DOI resolves to according to the location information that is registered for the DOI. The landing page has to be publicly accessible and contain up-to-date information, such as statements on how to access the data.

Additionally, there are some best practices in DOI management that are not requirements, but important to follow (for more detailed information, see [2]):

- In those cases when DOI-registered data become unavailable, with the consequence that the DOI resolves to an invalid, or non-existing page, the URL has to be updated to point to a persistent **tombstone page**. If the client cannot provide the tombstone page, the Allocation Agency can provide one.
- Data that are assigned DOIs should be on such **granularity level** that the data are easily and clearly citable.

According to **DOI syntax**, the clients are free to design their suffixes as they choose, provided the DOI is unique. The Allocation Agencies may recommend or provide guidelines for DOI syntax.

Cost

The DataCite services are paid for by the Member institutions, currently 17 full and 7 affiliated Members. Whether the individual Members, Allocation Agencies, will charge their clients or not for providing DOI prefixes depends on the business model of the Member organisation [2].

Quality of the Services:

DataCite uses the Handle System for DOI name resolution, and the world-wide HS resolver system is the backbone of the services.

If DataCite “is dissolved, reasonable steps are taken with the endeavour to maintain the resolution of DOI names registered by DataCite. This may include a request to IDF but shall include at minimum any steps necessitated by the contractual relationship with the IDF, if any” [1, p5], since DataCite is a Registration Agency of IDF (see section about Governance Structure).

User/Client Interaction

To start using DOIs via DataCite, presumptive clients should contact their local DataCite Member[10], who can provide them with access to the DataCite service for minting persistent identifiers (DOIs) and registering associated metadata.

The clients can contact DataCite through common channels (e-mail, Twitter etc). Their local DataCite Member can participate in working groups in which issues concerning the development of DataCite are discussed. The working groups interact with their clients on important matters.

DataCite organises a yearly Summer Meeting, open for all interested in PID development, and a General Assembly for the Member organisations.

Member organisations are supposed to create information material, organise seminars and arrange other promotional activities of their own.

User/Client Organisations

DataCite is primarily working with organisations that host data, such as data centres and libraries. See: http://www.datacite.org/members

Current Status

Most activity currently takes place in the Metadata Working Group, where Version 3 of the metadata schema is released, and a user forum is set up.

Best Practices Working Group initiated a survey among DataCite users and published a report.

Cooperation

DataCite are cooperating with several several organisations, projects and companies, some of the more interesting being CrossRef, CNRI, EPIC, ORCID, OpenAIRE, Thomson Reuter, and EuDat.

Usage within the DASISH communities

DataCite DOIs are used by 3 of the CESSDA archives – GESIS, UKDA, and SND – and usage is being implemented by the SHARE-ERIC.

References:

[1] DataCite, 2009. DataCite Statutes, Final Version 24 November 2009. Available at: <http://www.datacite.org/docs/datacite-statutes-final.pdf> [Accessed 2013-04-26]
The DOI system home page:
[2] DataCite, 2012. Business Model Principles, Version 1 2012□10□01. doi:10.5438/0007 Available at <http://datacite.org/sites/default/files/Business_Models_Principles_v1.0.pdf> [Accessed 2013-04-26]

Additional Links

http://www.doi.org
The DOI handbook:
http://www.doi.org/hb.html

DataCite:
http://www.datacite.org
Datacite-EPIC
http://www.doi.org/topics/EPIC_DataCite_March2012.pdf
http://www.doi.org/registration_agencies.html

---

[1] http://datacite.org

[2] http://datacite.org/TIB

[3] http://datacite.org/members

[4] http://datacite.org/board

[5] Data repositories, data centres, and organisations that hold data.

[6] http://www.doi.org/hb.html

[7] The resolution tool used in the DOI system is the Handle System(TM).

[8] http://www.doi.org/registration_agencies.html

[9] http://www.datacite.org/services

[10] Usually there is no more than one Member per country. If the clients' countries do not have any DataCite Member organisation, then they have to contact DataCite directly, who can inform them about which Member organisation to contact.

2.2. Towards a European Handle-based PID service – EPIC[6]

Introduction

EPIC was founded in 2009 by a consortium of European partners in order to provide PID Services for the European Research Community, based on the Handle System™, http://www.handle.net/), for the allocation and resolution of persistent identifiers. The consortium signed a Memorandum of Understanding aiming to provide long-term reliability for the PID services.

Mission Statement

---

[6] This description largely consists of a rearrangement of texts found on the EPIC website (http://www.pidconsortium.eu/), with some minor additions and clarifications. All of these parts have been approved by an EPIC representative. The only exception is the last sections "User experience", which is based on the survey mentioned there.

The goal of EPIC is to set up and maintain a reliable joint service for registering, storing, and resolving persistent identifiers based on Handles for the European research community. This is further specified under *Services* below. EPIC is open for use by any European institution that stores scientific/research data.

Organisation of the Handle System

The Corporation for National Research Initiatives (CNRI) designed, implemented, and currently administers the root level of the Handle System, but the bulk of the resolution services are managed by the thousands of organisations, communities, government agencies, and businesses around the world currently using the system on a daily basis. This includes many academic and national libraries, publishers of scholarly journals, scientific institution, and other information management groups. CNRI is working with other major Handle System user groups, including EPIC and DataCite, to create DONA (Digital Object Numbering Authority) to manage the Handle System in the future. The new organisation will be governed by the DONA Board, which will include experts and stakeholders from around the world.

Organisation of the EPIC service

The PID infrastructure used by EPIC is based on a worldwide two-level hierarchy. The Global Handle Registry and its global mirrors are on the highest level of the hierarchy. These systems are registries where the most important information of the prefixes is stored. Global Handle Registry Mirrors are therefore deployed on every continent. The GWDG, one of the EPIC founders, hosts a Global Handle Registry Mirror in Europe to ensure the resolution of the prefixes even if other parts of the global network are temporarily not available.

The EPIC consortium has enabled partner data centers to register any number of PIDs for the data objects and collections they store independently of their possible later use for publications. This registration in general will be done by automated procedures using a RESTful API that also allows adding relevant information such as checksums, pointers to metadata and rights information, etc.

In a Memorandum of Understanding, the following institutions declared the intention to start setting up and maintaining a joint service for registering, storing, and resolving persistent identifiers based on Handles.

- GWDG – Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen
- SURFsara – Reken- en Netwerkdiensten
- CSC – IT Center for Science Ltd.
- DKRZ – Deutsches Klimarechenzentrum

Governance Structure

EPIC is controlled by its scientific user communities and organisations to ensure that it is devoted to the needs of the research community at large. This will also ensure that the overhead of the EPIC consortium will be small and restricted to essential services.

Together with other stakeholders, the participating institutes will take part in and support the founding of an international governing board guiding further operation and development of the Handle System. The purpose of this is to safeguard the investments of the scientific community in using the Handle System for research data.

Services Offered by the EPIC System

EPIC sets up and maintains reliable joint services for registering, storing, and resolving persistent identifiers based on Handles for the European research community.

For users, the most important services are mainly the possibilities to generate PIDs and to be sure that these are resolved reliably and effectively for a long time to come. Here, EPIC provides the following services to the scientific research community:

- PID Service: Services to generate PIDs for digital objects
- Associate extra information records with the PID (e.g. checksum)
- PID Resolution: Services to guarantee reliable resolution of the PIDs, issued by EPIC
- PID Replication: EPIC replicates the databases of Handles to guarantee an robust and high-availability of the PID resolution function
- Global Handle Mirror Server: A mirror of the Global Handle Server in Europe.

For new service providers, the software stack of the PID service is made available, including documen-tation of the server software for all available versions. At the moment, this software is distributed without any support and with only a limited amount of documentation. Between members of the EPIC consortium, there is knowledge exchange and mutual support. Service providers that want to implement PID services under the umbrella of EPIC can ask for assistance on a voluntary basis at the contact address listed on the EPIC website.

Mandatory Requirements

Although EPIC recommends taking the persistency promise of PIDs seriously, it provides the tools to enforce any set of specific PID policies. For this, the EPIC service can be configured to follow precisely these policies inside a specific PID namespace.
This includes being able to ensure that e.g.

- PIDs should never be deleted (persistency of the identifier)
- PIDs should always contain a digital signature
- PIDs should follow a certain syntax

or other policy-specific requirements. Furthermore, EPIC will provide services that make the reliability of the PIDs of a namespace publicly transparent.

Costs

Running the Handle Services based on tested software is not expensive. However, EPIC sees the need to establish help and support services that will require some funding. Currently and for the coming years funding is ensured. Later on, contributions will be required. The User Board will determine the funding structure.

Quality of the Services

EPIC utilizes the Handle System to achieve a redundant and load-balanced setup between the data centers. EPIC replicates the PID databases to guarantee an all-time availability of the PID resolution. The integration with the global Handle infrastructure and the mutual mirroring of Handle Services between the EPIC partners guarantees a highly reliable and high-performance resolution service of EPIC-issued PIDs.

User Interfaces

The PID Service is the main interface for registering and managing persistent identifiers in EPIC. To help the users, the EPIC providers share the same interfaces. The PID Service is implemented as a RESTful web service and it is being continuously developed by EPIC.
A publicly accessible web interface for the resolution of Handles is available and integrated into the general and worldwide Handle framework for PID resolution.

A publicly accessible, EPIC-specific web interface for requests on Handles can be used to search for EPIC-issued Handles (the inverse of Handle resolution). If a modification of a given Handle's associated information or the allocation of a new Handle is required, the user has to log in to the system with their access data.

User Interaction

In order to enable allocation or modification of Handles, users and scientific institutions can apply for a test or real account at a central e-mail address to get access to the service. In the future, EPIC will provide a web sheet, which then needs to be confirmed additionally by a certified e-mail.

All new features and changes that are requested should be communicated to EPIC as change requests. EPIC registers these change requests, and a common decision about the implementation is made by the EPIC partners. If required, the PID service provider can set up a PID service on behalf of a scientific institution or community. Such a PID service will have its own PID prefix, which must be ordered at the CNRI registration form.

Users as well as providers can get in contact with EPIC via a central e-mail address listed on the EPIC website.

User Organisations

The participating institutions declare themselves willing to establish an appropriate sustainable service, operating- and business model that will extend the service already provided by the GWDG for the Max Planck Society. They will offer interested communities to participate in discussions about the principles of a shared, and therefore highly available and highly persistent service. During the first year, EPIC will work on a prototype solution for such a robust system with the intention to turn this into a full production service. It will enter into discussions with CNRI to find a proper basis for the smooth continuation of the Handle System and to establish the required independence. Other well-known institutions are welcome to participate in setting up and maintaining this shared persistent identifier system in Europe.

Cooperation

The following institutions and communities are currently supporting this initiative and will offer the services to its members:

- EUDAT
- Max Planck Society
- CLARIN
- DARIAH
- Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)

User Experience

In October 2012, a questionnaire was sent to CLARIN data centers ("Questions regarding the use of Persistent Identifier systems at the data center"). Several centers reported that they use EPIC:

- LINDAT-Clarin (CLARIN CZ).
- UdS Saarbrücken (CLARIN-D-Center)
- Meertens Instituut
- LMU Muenchen
- Hamburger Zentrum für Sprachkorpora (CLARIN-D-Center)
- University of Leipzig
- Universität Stuttgart
- CSC — IT Center for Science

Among other questions, the survey also asked for pros and cons of using EPIC. Among the advantages stated by the centers were the following points:

- CLARIN endorses the Handle System
- EUDAT also appears to move towards using Handles
- Quick response and support
- Good reliability
- Handle System is free and non-commercial
- Offers unlimited PIDs
- Allows for using testing and production prefixes.

The disadvantages which were mentioned can be summarized by these points:

- Already experienced non-responsiveness of service
- EPIC is as most services fast in check-out, but slow in changing PIDs
- API is still work in progress
- Provider responded slowly – still waiting for API availability
- Uses very long Handles
- Missing batch mode for PID registration
- Missing test instance of PID service for software components
- Not very suitable for citations.

The differences between EPIC and DataCite PID service usability can partly be explained by the different ambitions of service provider organizations. EPIC interprets persistency in a very wide

sense, foreseeing the need to provide PIDs for al type of objects or object-fragments, not only for citable data-collections, as is mainly the goal of DataCite. See [2]

Current Status

The current stable API v2.4 was developed by GWDG and SURFSara. It was released in May 2013 and has been productively used since then. The old API v1 will be deprecated in the first months of 2014.

Using the current API v2, there are over 20,000 PIDs from 20-50 institutional users (mainly CLARIN centers) stored with the GWDG, and around 6,000,000 PIDs (mainly from archives in the Netherlands) with SURFSara as of April 2012.

According to the GWDG, several improvements were made to the EPIC API as compared to the time of the above-mentioned questionnaire, when the available API was still work in progress. These improvements address some of the issues from the user responses and include:

- an improved responsiveness of the API, which EPIC achieved in close collaboration with CNRI, and which is still being developed
- an available batch operation mode
- determination of the Handle suffix by the user
- the user is free to chose which metadata to provide, in order to improve EPIC suitability for use in citations.

Furthermore, a workflow for transforming an EPIC-PID into a DOI is being worked upon. There is a process for ao. discussing improvements of management tools and API with user communities at regular conferences and workshops.

References

[1] EPIC website (http://www.pidconsortium.eu/) – last accessed 28/11/13

[2] EPIC DataCite goals (www.doi.org/topics/EPIC_DataCite_March2012.pdf - last accessed 2/1/13

2.3. Towards an European URN:NBN-based PID Service –the URN:NBN Cluster Service

Introduction

In contrast to the PID services DataCite and EPIC which came into existence on an international and European level the URN:NBN based PID services were developed on the national scale independent of each other. They are provided by the National Libraries and other national organisations (these could be other trusted institutions – for instance a national data centre like

DANS in the Netherlands). The first URN:NBN PID service has been introduced in Germany by the German National Library (DNB) in 2001. Since then, other European countries – such as the Netherlands, Sweden, Norway, Finland, Hungary, Italy, the Czech Republic, and Austria – have set up their own URN:NBN services. Currently not all EU countries are covered.

The URN:NBN service providers are acting independent of each other and each one has its own policy. The disadvantage of this infrastructure is that it leads to many different URN namespaces that are not interoperable. Furthermore the variety of policies on allocation prohibits transparency. To overcome this situation the PersID project came into existence. It persisted from October 2009 until March 2011. The goal was to harmonize and network the different European PID service solutions and to initiate a global governance infrastructure.

With the new URN:NBN Cluster project the PersID initiative and its approach will be continued. It started in November 2011 and there are on-going work to create a single point of entry resolving URN:NBNs from any namespace, including PIDs from other trusted PID services like Handle and DOI.

Mission Statement

The URN:NBN Cluster has been established under the consideration that PIDs based on URN:NBNs are based on internet standards, are openly available, and not owned by an organisation or vendor. The cluster will be totally discipline-independent.

Organisation of URN:NBN PIDs and the national URN:NBN services

In the PersID project the chosen PID systems are URNs, Uniform Resource Names. The URN system is an Internet standard governed by the Internet Engineering Task Force (IETF). URNs are intended “to serve as persistent, location-independent resource identifier”.

The URN scheme is composed of a prefix and a suffix consisting of namespaces (urn:[Namespace Identifier]:[Subnamespace Identifier]-[Namespace Specific String]. NBN is a registered UNR namespace Identifier and stands for National Bibliographic Number. Usually National libraries administrate them. They may be assigned to a wide variety of digital objects. Other organizations may obtain a sub-namespace via their national library and assign identifiers independently, on condition that they must adhere to the national NBN policy.

An important aspect of the URN:NBN PID services is that no direct metadata search is being offered. In the vision of the URN:NBN Cluster, a PID service will lead a user either to a metadata description of a resource (a landing page) or directly to the actual resource.

Separate URN:NBNs will be assigned to different versions of an object. In other situations (not stable intermediate versions) the URN:NBN will only be assigned to the final version. No URN:NBNs will be assigned to dynamic datasets. In the URN:NBN philosophy, the goal of URN:NBN identifiers is to make digital objects permanently citable and accessible, and therefore these objects should be preserved in an unchanged state. A dynamic dataset does not meet this criterion.

When a namespace has been assigned to a particular organisation, usually a National Library, (by the Library of Congress), this organisation is entitled to assign URN:NBNs to digital objects under the following conditions:

1. The URN:NBN will be persistently accessible
2. The registered digital resource corresponding to a certain URN:NBN will be preserved persistently
3. The registered digital resource corresponding to a certain URN:NBN will be persistently accessible, but additional conditions may apply (such as authorisation or copyright) for obtaining the resource.
4. A URN:NBN may be used only once
5. A URN:NBN may not be re-used
6. URN:NBNs only become valid once they and their associated URLs have been registered at a National Library
7. All the different representations of an object will have the same URN:NBN
8. If the content of objects has changed, then a new URN must be assigned
9. Making URN:NBNs actionable as URLs is necessary to access to the global resolver. Example: urn:nbn:nl:ui:13-o4p-8py may become:

   http://www.persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-o4p-8py.

Organisation the URN:NBN Cluster

PersID was a cooperation project between ten national organisations in eight European countries. In November 2011, the DNB, DANS, and the National Library of Sweden agreed on continuing the ideas of PersID and the development of the URN:NBN Cluster project.
In 2012, DNB started the implementation of the URN:NBN Cluster. Attention is also being paid to the legal aspects of the cooperation. The intention is to set up a legal body to assure its maintenance and operation. In November 2013, all participating institutions (the National Library of Sweden, DANS, and DNB) have signed a Letter of Intent. The aim now is to use the experience gained from the prototype to eliminate errors, to simplify the implementation, and to improve the documentation. This will provide the basis for gaining additional partners for the cluster, but currently there is no EU wide coverage.

In the cluster, all partners are sharing responsibility. The functionality offered by the cluster are determined by the partners and based on what is described in the PersID project. By setting up a network of connected (national) Namespace Resolving Services with a central Global Resolving Service on top of this, the service is quite robust, all the more so as these Namespace Resolving Services mirror each other.

Services Offered by the URN:NBN Cluster

The planned service offered by the URN:NBN Cluster will provide a common infrastructure for URN resolving in 2014, to overcome the situation that each namespace has its own resolver. However currently this is not yet available.

Costs

The *business model* of the cluster is based on contributions from the partners. As these are, in general, governmental organisations, the national European governments contribute indirectly. Currently a detailed business, operational, and support model for the URN:NBN Cluster is developed.
Apart from this, the national URN:NBN services are paid by organisations working on a national scale, like the National Libraries. So indirectly, the costs are covered by the national governments.

Current Status

DNB has started a broad URN:NBN-based service (in the beginning only for Germany, Austria and Switzerland). At the moment the German National Library finished the proof-of-concept phase together with the Swedish National Library and support from colleagues from DANS (NL). Now the prototype is transferred into the productive phase in Sweden, the Netherlands, and Germany. The software for setting up a meta-resolver is already available via SourceForge: http://sourceforge.net/projects/metaresolver/. Moreover, Austria has established its own URN:NBN system, based on the software developed by DNB. In this way, a community has been formed.

The goal of the DNB and its partners is to create a common infrastructure for URN:NBN resolving in Europe, with one resolving service as a single point of entry with high availability for the multiple (national) URN namespaces. Administration and resolving of URN:NBNs will be separated in this infrastructure.

References

Policy for Issuing URNs in the URN:NBN:DE Namespace, version 1.0, November 2012, http://d-nb.info/1045320641/34 (last accessed December 12th, 2013).

Developing an International URN:NBN Cluster (2013). Presentation from U. Ackermann, Jan Hannemann, N. von der Hude, and Th. Siedel given December 2012 on the Conference on Interoperability of Persistent Identifiers Systems, held in Firenze, Italy. http://www.rinascimento-digitale.it/conference2012/w_pi_2012/dbn_cluster.pdf (last accessed December 12th, 2013).

Der Uniforme Resource Name (URN), Version 2.0. Chr. Schönung-Walter. In: Nestor Handbuch: eine kleine Enzyklopädie der digitalen Langzeitarchivierung (2009), Kapitel 9.4.1. urn:nbn:de:0008-20090811490 (last accessed December 12th, 2013).

2.4 Comparison of the PID Services

	DataCite	EPIC	URN:NBN Cluster
business model	Financed by member institutions. Member institutions may charge clients.	Currently maintained by MoU partners. Future funding models tbd by User Board	Contributions from organisations active on a national scale
governance structure	General assembly consists of member institutions.	controlled by its scientific user communities and organisations	Network of organisations with shared responsibilities
functionality	Extended services based on the Handles system.	Extended services based on the Handles system.	Global Resolving Service,GRS, communicating with local Namespace resolvers
robustness	Part of the IDF (International DOI Foundation) global network of resolvers.	primary LHS at each MoU partner, each primary LHS mirrored by all other partners; EPIC runs the only Handle proxy outside of US	A Network of Namespace Resolving Services creates a robust and stable infrastructure with a 24/7 availability
availability	High availability	has been criticized, recently improved, still in development	see: robustness
coverage	World-wide	EU-wide	Limited to some EU countries
availability	High availability	has been criticized, recently improved, still in development	see: robustness Not available EU wide
scalability	High scalability	depends on size of consortium	Has to be addressed yet
persistency	If an object is not available a tombstone page is mandatory.	Depends on user selected namespace policy	Realised via shared responsibilities

tightly coupled metadata e.g. object checksum	In principle, but not used	yes	no

3. Results and Presentations of the Questionnaires

In order to assess the different user requirements for PID service providers mentioned in the Introduction, DASISH T5.2 sent out two different questionnaires: one to the five DASISH communities within the social sciences and humanities (CESSDA, CLARIN, DARIAH, ESS, and SHARE), and another to the individual data centres within these five communities.[7] The second questionnaire was designed to get more detailed opinions of the immediate clients (i.e. data centres and archives) and potential clients within these communities on benefits, drawbacks, and desired improvements in using certain PID services. The results from both these questionnaires are summarized in the following two subsections.

ESS and SHARE are two major European survey programmes and not data centres. Each of them outsources the archiving of their data to external data centres (ESS to Norwegian Social Science Data Services, and SHARE to CentERdata and GESIS). Consequently, for those two communities it is only possible to get a general answer of the community.

3.1. Responses from the Communities

The goal for this part of the questionnaire was to get an overview of already available policies/routines regarding PID services at the community level, to determine in what percentage of the data centres within these communities that PID services are being used, and to get a picture of the crucial requirements at the community level.

Question 1: Are there any policies/routines for PID services at the community level?

Figure 1

	Yes	No (but in progress)	No
CESSDA			✓

[7] Please find the two questionnaires in the appendix.

CLARIN	✓		
DARIAH			✓
ESS			✓
SHARE		✓	

Up to now only CLARIN has produced an internal policy document on Persistent Identifier in 2009.[8] This document recommends the Handle System for PID services. The idea is that all potential CLARIN centres will use it. Where institutions do not run their own Handle System resolver, CLARIN recommends using EPIC as a PID service provider. This recommendation was partly based on the old DOI business model (pay per assigned DOI). Meanwhile, the different DOI services within DataCite have their own business models[9] and some of them are at no charge[10] (e.g. DataCite Germany). Both EPIC and DOI are interoperable, using the same underlying HS technology, and there seems to be an emerging consensus that DOI should be used for published data sets (requiring only a limited number of PIDs), while EPIC services are better suited for managing very many PIDs, for instance in scientific data flows and accessing individual resources.

SHARE is about to register with DOIs. They are in contact with the da|ra registration agency for social science and economic data[11] and already received a prefix for their DOI name. The registration process is still in progress. Regarding the other communities, it appears that within ESS, DARIAH and CESSDA no policies on PID services have been implemented yet.

Question 2: Approximate how many/what percentage of the data centres in the community use PID services.

[8] http://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf
[9] http://www.datacite.org/sites/default/files/Business_Models_Principles_v1.0.pdf
[10] http://www.tib-hannover.de/en/the-tib/news/news/id/362/
[11] da|ra is the registration agency for social science and economic data jointly run by two of the four German representatives in the international DataCite consortium GESIS – Leibniz Institute for the Social Sciences and ZBW - German National Library of Economics and Leibniz Information Centre for Economics. da|ra offers its service at no charge in conjunction with DataCite to further sciences beyond social sciences and economics and moreover to institutes outside of Germany. More information on http://www.da-ra.de/en/home/

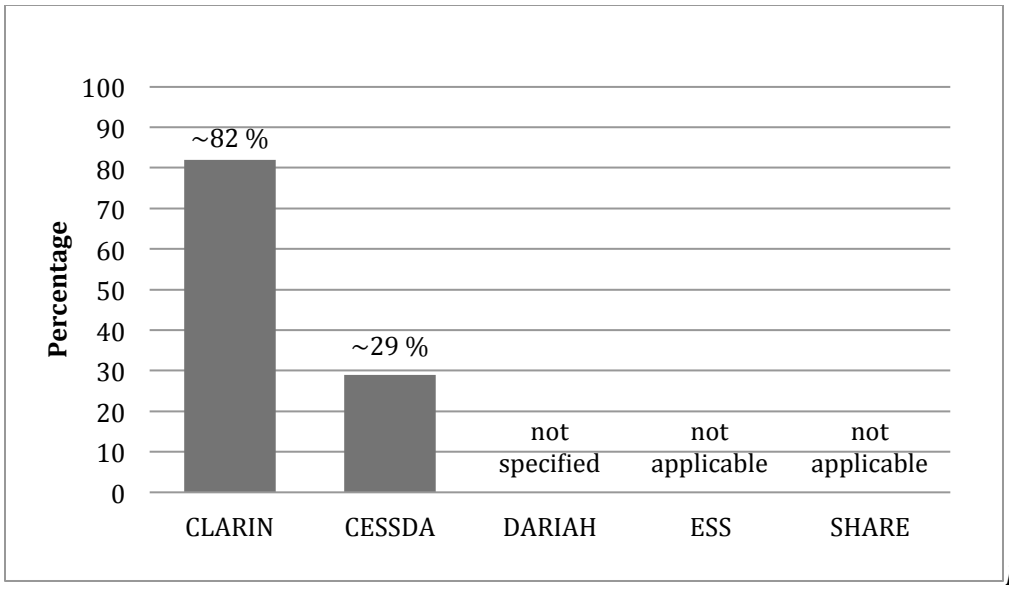

Figure 2

Within the CLARIN community approximately 82 % of the data centres are using PIDs and PID services.[12] Nine of them are using Handles via EPIC, three Handles via their own server and two URN:NBN. The use of PID services is under construction at the Clarin.dk centre and will likely be ready in 2013.

The estimation for the CESSDA community is 29 % (6 in 21 centres). Four of the centres are using DOIs and two are using URN:NBN.

For DARIAH, no estimation is available. But the result of the analysis of the second questionnaire was that at least six centres are using different PIDs and services (ARK, DOI, Handle, Crossref, EPIC, and a PID used by the SUDOC reference registry).

Because SHARE and ESS are surveys consisting of several research centres but not of data holding institutions question 2 is not applicable for these two DASISH communities.

Question 3: Are there any additional requirements on the PID services at the community level?

In 2009, CLARIN also published requirements for PID services.[13] The other communities have not produced such a document yet.

Most important requirements mentioned in the CLARIN document are:

- A PID registration and a robust, reliable and persistent resolution service with scalable architecture, fast hardware and network connection that is available 24/7, and long-term supported from governments.
- Association of a PID with the original object.

[12] https://centerregistry-clarin.esc.rzg.mpg.de/
[13] http://www.clarin.eu/files/wg2-2-pid-doc-v4.pdf

- A PID -service open to all disciplines.
- Explicit rules for PID policy on version and fragment identifiers.
- A PID syntax complying with the IETF standards.
- A PID service offering limited descriptive metadata.
- The PID syntax and the resolution mechanism must accept the usage of fragment identifiers.
- A high security level for the resolution database and a regular backup.
- Independence/openness of the resolution software (free of constraining licenses).
- A PID service business model not linked to (or dependent of) the number of PIDs and resources.

3.2. Responses from the Data Centres within the Communities

The aim of this questionnaire was to gather information on the use of PID services at the centres of CESSDA, CLARIN and DARIAH, as well as on their PID policies, purposes and requirements. Furthermore, the centres were asked about the obstacles preventing them from using PID services and the services helping them to start with it.

Overall, 28 responses were submitted and 26 could be evaluated. At the time of the survey CESSDA had 21 and CLARIN 20 members. PIDs are an important subject for these institutions as they all have an archiving function. The structure of DARIAH is different. On the one hand, it has more member institutions – at the time of the survey 67 – but on the other hand, most of its centres have no archiving function and it was not possible to detect the exact number of the data holding institutions. The questionnaire was sent to all CESSDA and CLARIN members and 10 DARIAH members who were rated as archives.

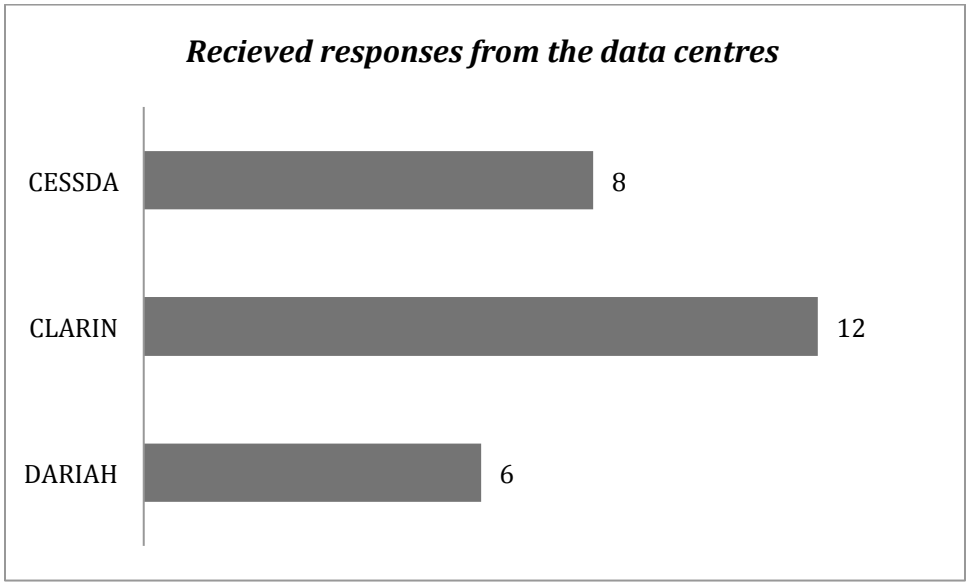

Figure 3

Question 1: Are there any policies and/or routines for the use of PID services at the data centre?

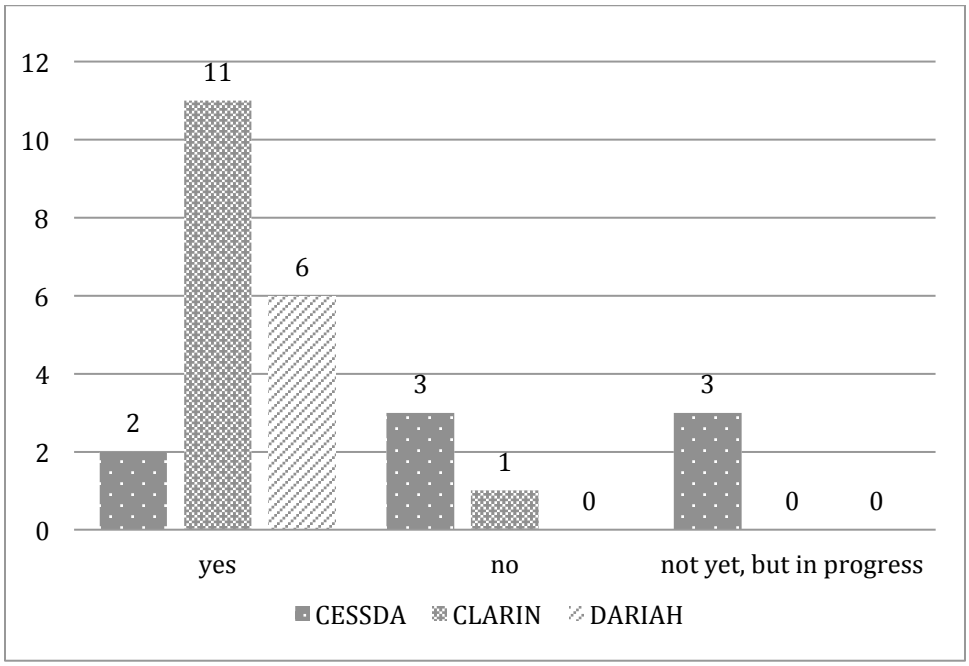

Figure 4

Question 2: Does the data centre use PID services?

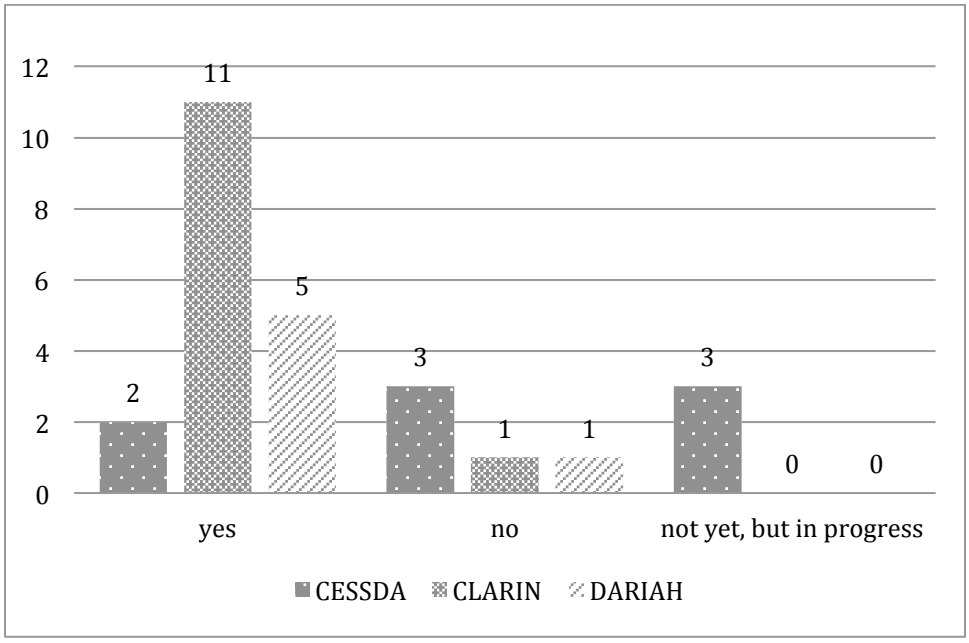

Figure 5

Question 3: If the data centre uses PID services: why does the data centre use PID services? What purposes and expectations/requirements does the centre have?

For this question participants had to select the requirements from a list. Multiple answers were permitted and requirements which were not mentioned could be listed below (= other specifications). For 4 in 26 centres the question was not applicable. 2 in 26 responses could not be evaluated. The first schema shows how often the requirements were checked/not checked by all DASISH respondents. The second one shows the distribution amongst the communities.

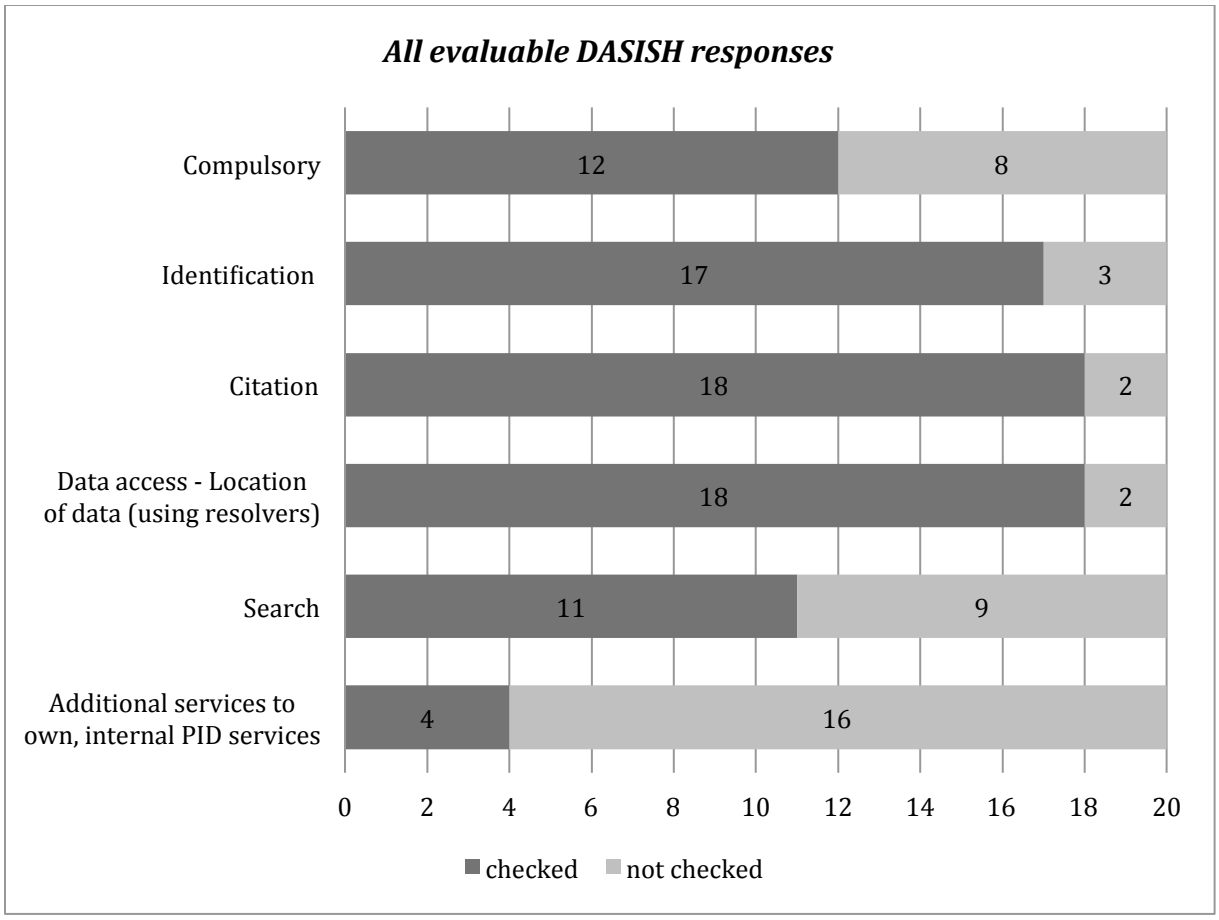
All evaluable DASISH responses
Compulsory
Identification
Citation
Data access - Location of data (using resolvers)
Search
Additional services to own, internal PID services
12
8
17
3
18
2
18
2
11
9
4
16
0
2
4
6
8
10
12
14
16
18
20
checked
not checked

Figure 6

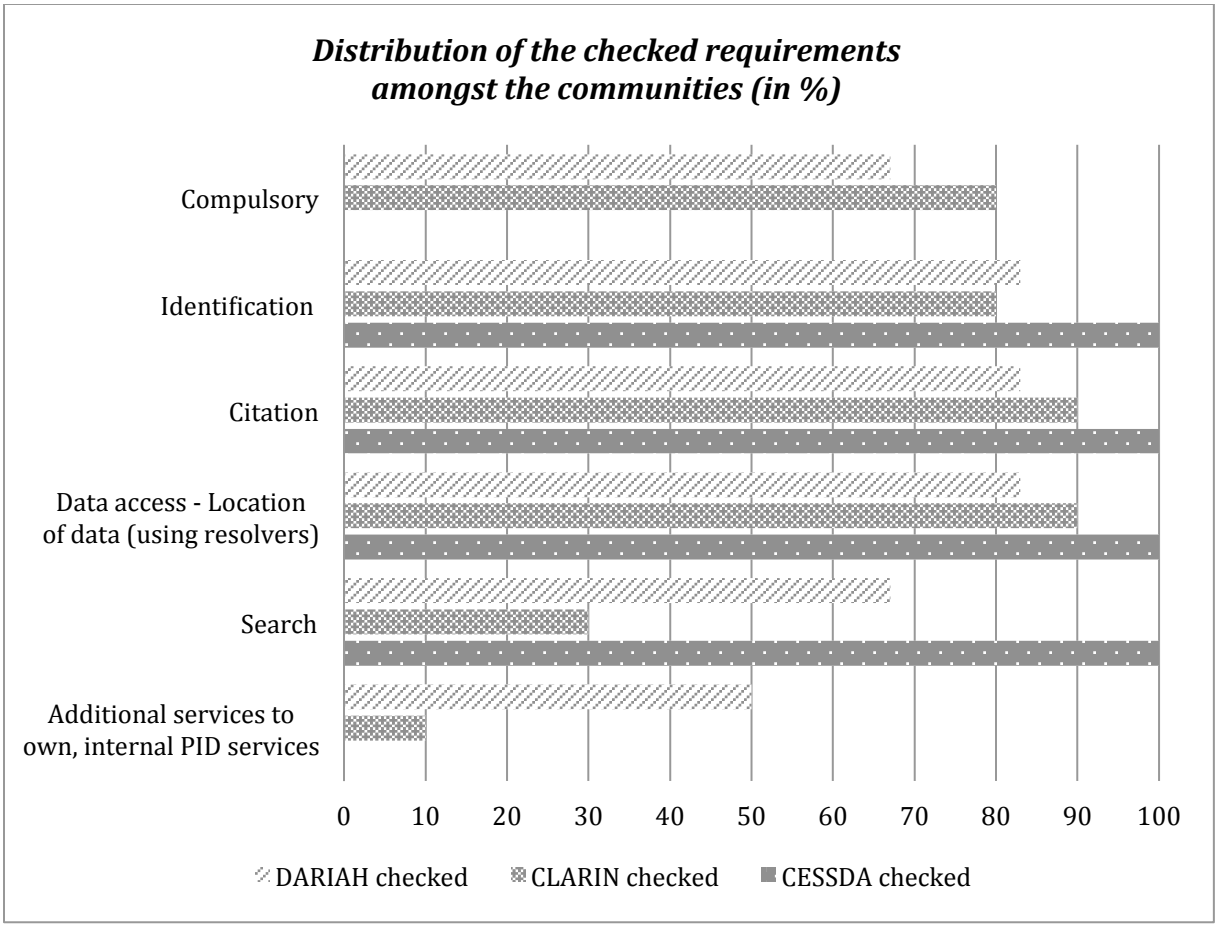

Figure 7

Other specifications:

- Clear identification of versions.
- A new version of data is linked to the previous version using their PID.
- A Data Citation Index.
- Localization of and access to cited data should be facilitated through machine actionable persistent identification.
- Traceability of research and efforts with regard to link literature, data, and authors.
- Feeding a harvester via OAI-PMH.
- To "objectify" data by relating metadata to data via PID.
- Access to numerous data cross-referencing other sources.
- Use of PIDs for direct access to specific formats (RDF, Unimarc, XML).
- The European Bioinformatics Institute (EMBL-EBI), with staff of 430 and based in Hinxton, Cambridge, uses persistent identifiers. The persistent identifier systems have been adopted widely by the biomedical communities EMBL-EBI serves over many years. For example, PDB Identifiers for protein structures, Accession Numbers for nucleotide databases, PMIDs. Literature databases store DOIs but do not use them as the principal

identifier system, using PMIDs and PMCIDs as the key identifiers. PDB issues a DOI as well as a PDB ID for each record, but this was a decision of that particular database.

- The ELIXIR nodes have worked with the biomedical community widely over the years to have PID's available. After ELIXIR becomes operational, there could be a possibility to define a standard. For the Finnish ELIXIR node the implementation of PIDs will follow BBMRI.fi choices (biobanking community) along with the ELIXIR/EMBL-EBI.

Question 4: If the data centre does not use PID services: what is/are the main reason/s?

For these questions, participants had to select reasons from a list. Multiple answers were permitted and reasons which were not mentioned could be listed below (= other specifications). For 4 in 26 centres the question was applicable. Three responses are from CESSDA, one is from CLARIN.

a) If interested in using PID services; are there any obvious obstacles that prevent the data centre from using PID services?

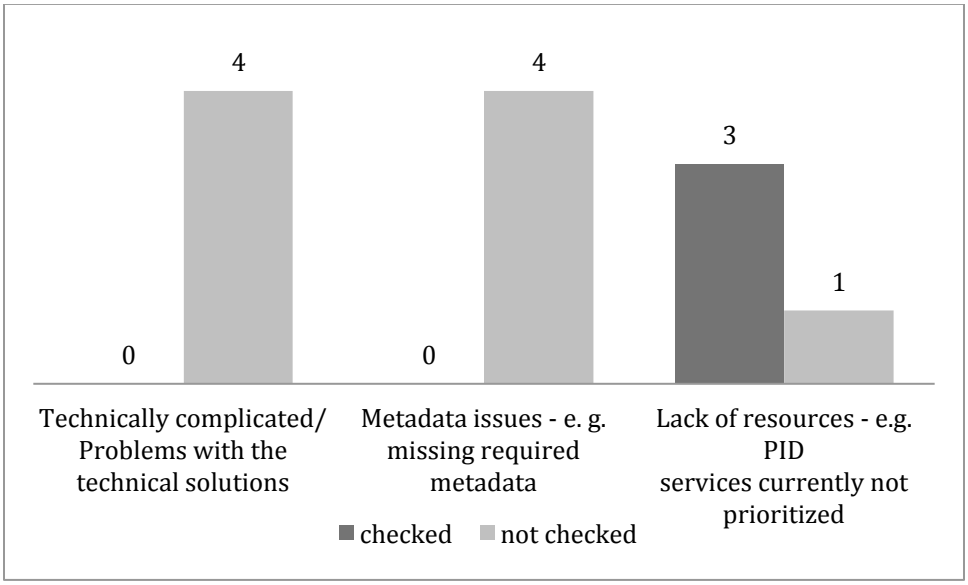

Figure 8

Other specifications:

- Every dataset gets a unique ID when deposited in the archive of the centre. Under normal circumstances that ID will not be changed later. For now, the centre considers this practice sufficient, but is interested in PID services and sees the advantages of such a service. Currently there are national and international projects that investigate PIDs, and it is its decision to wait and see the emerging solutions and recommendations. Therefore, one might also say that PID services are currently not prioritized.

b) If not interested in using PID services, why?

All evaluated centres are interested in using PID services.

Question 5: If the data centre does not use PID services: what would it take for the data centre to start using PID services? Actions taken by the PID service providers based on requirements (specification) from the Communities.

For these questions, participants had to select actions from a list. Multiple answers were permitted and actions which were not mentioned could be listed below (= other specifications). For 4 in 26 centres the question was applicable. Three responses are from CESSDA, one is from CLARIN.

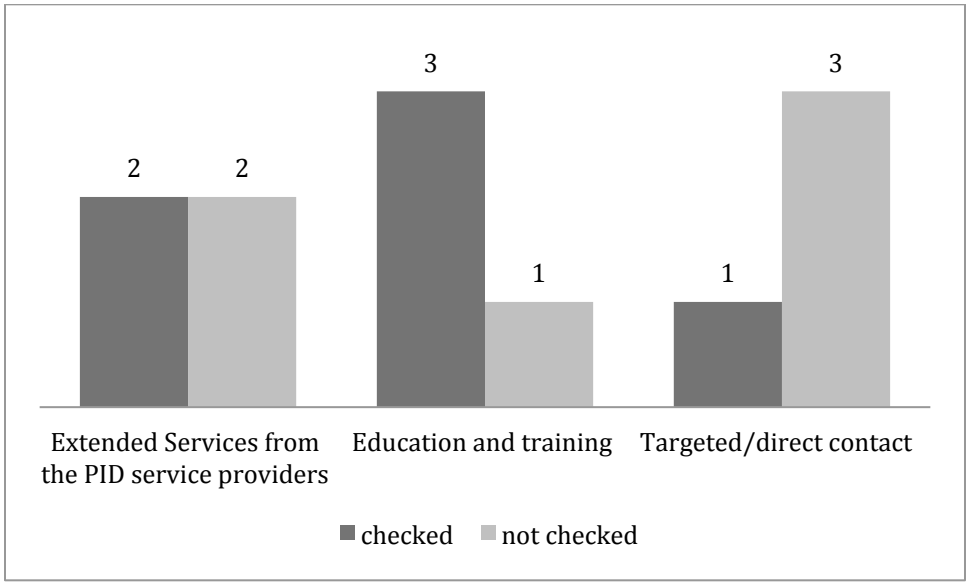

Figure 9

Other specifications:

- In order to start using any particular PID services, the centre would need training and, depending on the nature and conditions of the PID service, some additional funding.

Question 6: What are the pros and cons with those PID Services that are currently recommended/used by the community/centre (experiences on both overall and detailed levels)?

CESSDA

Specifications of centres using DOIs/DataCite:

- extended metadata, e.g. used in search for both data and other publications simultaneously
- widely used and accepted
- complete infrastructure for DOI registration and metadata administration

- extensive metadata description-schema including domain-specific items
- search in metadata
- metadata transfer is possible in three different ways: web interface, xml upload, and web service
- free of charge
- good reasons to use DOI names: good prospects of dispersion and persistency and a supervising organization of the IDF
- membership of DataCite: internationally coordinated approaches
- future: offers regarding impact factors and peer review and the linking of the data and publications.

Specifications of centres using URN:NBN:

- works fine and distinguishes itself from other PID systems in the high trust level: institutions adhering to the URN:NBN system commit themselves not just to maintaining the PIDs, but also the PID infrastructure (registration agency, resolver) and the resources identified by the URN:NBN
- the URN:NBN system currently lacks a widely approved syntax for identifying fragments; identifying fragments, for instance, paragraphs in a document, is useful for referring directly to specific data rather than citing a whole dataset
- an organization saving data/materials that are related to PID needs to ensure that servers will be available; so there are additional finances related to this – new hardware/software.

Specifications of centres not using PIDs:

- commercial vs. public funding
- dependency on (current) technology
- level of standardization
- ”human-understandability”
- expandability
- user communities
- aspects of long-term preservation
- version control
- citing data, a tiny thing that is going to dramatically change how scientists perform their research.

CLARIN

Specifications of centres using Handles/EPIC:

Pros

- experiences with PID provider(SARA/EPIC) are very good: quick response and support during implementation change
- good reliability

- reasonable use (technically)
- free
- unlimited PIDs
- “testing” prefix as well as a production one
- CLARIN endorses the handle system as it seems to be the only one that fulfils all requirements and the centre follows this recommendation
- all CLARIN D-Centres use Handles
- the centre uses Handles from GWDG and has currently no complaints about them

Cons

- strong dependency on stability of PID services (single point of failure, occurred briefly twice in two years)
- very long Handles, not very suitable for citations
- GWDG was rather slow in providing the centre with an account for using the Handle system; everybody is still waiting for the API V2 to be made available
- API is still a work in progress
- the PIDs themselves are fine for our usage scenario but the centre is missing a test instance of the PID service that allows them to test software components that work with a PID system (e.g. register PIDs) without making changes to the productive version of the system. They would also like to have a way to register a large number of PIDs in some kind of batch mode (one request to register 100 PIDs instead of 100 requests).

Specifications of centres using URN:NBN:[14]

- works fine and distinguishes itself from other PID systems in the high trust level: institutions adhering to the URN:NBN system commit themselves not just to maintaining the PIDs, but also the PID infrastructure (registration agency, resolver) and the resources identified by the URN:NBN
- the URN:NBN system currently lacks a widely approved syntax for identifying fragments; identifying fragments, for instance, paragraphs in a document, is useful for referring directly to specific data rather than citing a whole dataset.

Other specifications:

- the centre would prefer to support non-commercial systems
- the check-out of a new PID must be fast, since the centre checks out/changes millions of PIDs; most services are reasonable fast in check-out but too slow in change
- the PID service should provide additional, structured “verbose” information attached to the PID; important for the centre are: URL to metadata, author, date
- availability: the centre already experienced “blackouts” (PID service was non-responsive)

[14] The following comments are listed once more at this point because the centre is member of two communities.

- The centre didn't investigate other services as it is compulsory for CLARIN to use Handles; they are investigating the move to EPIC by SARA
- the centre uses Handles from GWDG and did not investigate the alternatives

DARIAH

Specifications of centres using ARKs:

Pros

- ARK is a free persistent identifier standard
- the ARK identifier is a reference for a unique resource, which in turn can have multiple identifiers; the persistency doesn't rely on the physical location of the resource, and is ensured by a Name Assigning Authority
- several others functionalities are interesting, such as the identification on different levels, allowing access to different versions of a same resource, allowing the creation of links between different objects
- the documents can potentially be accessed by a browser, using the address toolbar and a resolver; this can be useful for citations via the Internet, and allows the use of bookmarks
- a hierarchical relationship between objects can be implemented using the "/" (slash) character in the name section of an ARK; this part is optional and not persistent; it could be associated to services that can change or disappear; so, a distinction can be done between the identifier ark:/12148/bpt6k85329c representing the logical object and the identifier http://catalogue.bnf.fr/ark:/12148/bpt6k85329c/f4.pagination which represents the physical object corresponding to the 4th page of the document; only the former is persistent

Cons

- no existing off-the-shelf PID resolver which can be deployed as-is (in-house development to be planned)
- other existing PID standards (e.g. Handle, DOI …) in the community; need to limit the number of PID systems

Specifications of centres using DOIs/crossref:

Pros

- strong incentive for improving metadata quality of the platforms

Cons

- dependence on an external agent
- DOI deposit costs
- limited services compared to the lack of support and the constraints imposed by the service provider
- some technical limitations (lack of usability for tracking deposit errors)

Specifications of centres using Handles:

Pros

- light weighted
- inexpensive
- easy-to-use infrastructure to provide PIDs independently of the underlying implementation

Cons

- lack of uniformity (DOIs, ARK, Handle …)
- a great many different providers even at a national level

Other specifications:

- there are only advantages in using PIDs; nevertheless, the rights associated to open data (with a PID access) and their possible use has to be considered; the centre is about to finalize license(s) concerning open data provided online

Question 7: Other comments and additional requirements regarding PID services?

CESSDA

- It needs to be kept in mind that PID services alone are not enough to assure permanent access to data or metadata; there needs to be a reliable, long-term national/international service provider to maintain the systems (most notably a resolver and registry) and to offer services and support.
- Other issues to consider include the granularity and version management, and what a PID really does identify (metadata, data, jump-page?).
- How to deal with eventual multiple PID services (and PIDs)? Are these interoperable?
- Make various existing PID systems interoperable; the associated PID service providers should collaborate on this because interoperability will stimulate researchers' willingness to use PIDs and make collaboration within and among communities more efficient; a relevant report by APARSEN, which also presents the findings from a user survey, is available at http://www.alliancepermanentaccess.org/wp-content/plugins/downloadmonitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework.

CLARIN

- Although the centre runs its own PID service and does not have any serious issues with it, they decided that if a stable and reliable external PID service becomes available, they would switch to using that, because having their own service does not provide any additional value; however, they do not think such a stable and reliable service yet exists.
- Additional requirements: persistency, redundancy, scalability, embedding in European network, and a strong user platform.

- Current resolvers are slow and unstable, and don't really provide any more persistence than DNS – and are actually dependent on DNS, e.g., to perpetually resolve hdl.handle.net and dx.doi.org.
- Services that require a specific kind of PID for digital objects are undesirable, because in order to use many such services, digital objects need to have multiple PIDs; this also means that there is a great deal of duplicated work when all PID using parties need to join multiple PID consortia in order to get the PIDs.
- How to resolve the URL of the PID service from the pure PID? For instance the PID 11858/00-1779-0000-0007-CF8F-8 was issued by the handle system at GWDG (coded in the prefix "11858"); the naïve user does not know this, nor does a web browser.
- How can PID prefixes of different PID services resolved to the URLs of the handle system (here: http://hdl.handle.net/)?
- The practices of the centre regarding PID registration might change in the future in order to adhere to best practices.

DARIAH

- The data centre is considering future use of external PIDs like DOI or ARK.
- A robust resolution technology is mandatory; the business model has to be controlled by the scientific community.
- A process for creating PIDs of documents pointing at their latest versions in addition to the version-specific ones should be figured out; this should be done in collaboration with partner data centres for the sake of interoperability.
- Currently the centre stores the ARK identifiers of archived items in the DESC field of their handles; the centre hopes that a process will be devised to create a stronger link between PIDs and ARK identifiers.
- Need for a meta-resolver at a national level.

3.3. Conclusion and Analysis

Except for CLARIN, no policies regarding PIDs and PID services have been developed so far within the DASISH communities. At first sight, this may seem to be a major problem, but on the other hand, this fact offers an opportunity for rapid introduction of policies based on the findings within CLARIN and the findings of earlier European reports (see appendix 3).

Not all data centres in the communities are using PID services and it has to be stressed that even within the CLARIN community not all data centres are currently using the same PID system. Although new requirements have been set up that require the use of PIDs based on the Handle System technology. The same is true within the CESSDA community. CESSDA is an old community (1976) and most of the member archives have their own way of securing permanent access to their holdings, and ready-made citation statements. Therefore it is necessary to promote not only the use of PID systems in general but also to consider the interoperability between PID systems as an alternative to promoting (or even prescribing) the use of a particular PID system. Work in this direction has, for instance, resulted in the "Den Haag Manifesto" which describes some options to interoperability.[1]

Important purposes and requirements for the DASISH centres are "citation", "identification", and "data access". "Search" and locate data seems to be more relevant for CESSDA and DARIAH centres. "Compulsory" and "additional services to own, internal PID service" are only applicable for CLARIN and DARIAH centres. Other specifications are versioning, data citation index, relating metadata, cross-referencing, and broadening the concept of persistent identifiers to authors (and organisations) in order to realise persistent relationships between objects and their creators (within their work environment).

All evaluated centres are interested in using PID services. The main reason for not using PIDs and PID services is a "lack of resources" i.e. "not prioritized", and the decision is to wait for emerging solutions or guidelines from the community. These centres are in need of "education and training" and/or "extended services from the PID service providers". These findings will be communicated to the SSH communities who in their turn can support their data centres in start using PID systems. They will also result in best-practise guidelines and other educational material in cooperation with WP7.

The experiences with EPIC are good. Improvements that the centres would like to see include better stability and more user service of the PID service, a test instance of the PID service and the possibility to register large numbers of PIDs in some kind of batch mode. These issues have been addressed by the recent improvements in the EPIC service (see section 2.2).

The URN:NBN service works fine on a national level, not yet available on a pan- European level,(see chapter 2.3) and supports both the commitment to PIDs and to the PID infrastructure (registration agency and resolver). A major disadvantage is the absence of a widely approved system for identifying fragments.

The specifications on DataCite are very positive. The extended metadata and the search in metadata are emphasised in particular.

Other comments worth observing include a service free of charge, long-term preservation, version control and granularity, additional information attached to the PID (URL to metadata, author etc.), availability and permanent access, lack of uniformity of the different PID systems and providers, interoperability of the different PID systems, a user platform, and a resolver for all PID systems.

The results of the survey can be understood and summarized as the following list of DASISH requirements:

1. DASISH requirements for a basic PID service

(Requirements reduced to the least common denominator within DASISH)

1. A PID registration and resolution service infrastructure has to be available under the responsibility of a reliable and long-term funded organisation, operating at a European or national scale.
2. It maintains the systems and offers services and support and is embedded in a European/national network.
3. It has a clear policy that describes the responsibilities of the different stakeholders.
4. It offers a minimum set of descriptive metadata: e.g. title, author, publisher, publication year, rights, PID.
5. The resolution technology has to be reliable, fast, persistent and scalable and is 24/7 available.
6. The business model has to be sustainable for all involved stakeholders. It is controlled by the scientific community.
7. It is available to centers and users EU wide.

2. DASISH requirements for an extended PID service

(Requirements for a future scenario represented by particular groups within DASISH)

1. The PID syntax and resolution mechanism of the PID service must accept the usage of version and fragment identifiers. The PID service provides support for the version and fragment management.
2. The PID service supports the traceability of research and efforts with regard to link literature, data and authors.
3. It provides different representations/formats of metadata associated with PIDs (content negotiation), and can ideally be assigned to authors and organisations.
4. The rights of an individual PID is owned by the author/organisation that produced the object to which the PID has been assigned.

[1] http://www.knowledge-exchange.info/Default.aspx?ID=462

3. DASISH requirements for extra services

1. PID services within the ESFRIs have to be interoperable. Users should not be confronted with the PID diversity. To be a able to resolve all types of PIDs there should be a meta-resolver service that allows users to enter any type of PID and resolve it.
2. Education and Service for the data centres regarding PIDs in general are needed.

4. Comparison of the PID service providers in the light of the DASISH requirements

These requirements and the three PID services are compared in the table below. Requirements for which there is a need for action are shown in the table.

A Overall DASISH requirements regarding PIDs and PID services	DataCite	EPIC	URN:NBN Cluster
1. Interoperability	x	x	x
2. Education/Training	no	(x)	x

B DASISH requirements for a basic PID service (Requirements reduced to the least common denominator within DASISH)	DataCite	EPIC	URN:NBN Cluster
1. A PID registration and resolution service infrastructure has to be available under the responsibility of a reliable and long-term funded organisation, operating at a European or national scale.	x	x	x
2. It maintains the systems and offers services and support and is embedded in a European/national network.	x	x	x
3. It has a clear policy that describes the responsibilities of the different stakeholders.	x	?	x
4. It offers a minimum set of descriptive metadata: e.g. title, author, publisher, publication year, rights, PID.	x	no	no
5. The resolution technology has to be reliable, fast, persistent and scalable and is 24/7 available.	x	x	x
6. It is available for centres and users EU-wide	x	x	no
7. The business model has to be sustainable for all involved stakeholders. It is controlled by the scientific community.	x	x	x

C DASISH requirements for an extended PID service (Requirements for a future scenario represented by particular groups within DASISH)	DataCite	EPIC	URN:NBN Cluster
1. The PID syntax and resolution mechanism of the PID service must accept the usage of version and fragment identifiers. The PID service provides support for the version and fragment management.	x	x	no
2. The PID service supports the traceability of research and efforts with regard to link literature, data and authors.	x	?	no
3. It provides different representations/formats of metadata associated with PIDs (content negotiation).	x	x	x
4. It offers a technical test instance to test software components without making changes to the productive version.	x	x	x
5. Objects to which PIDs can be assigned may also be authors and organisations.	no	no	no

The PID services analysed in this task are all suited for use within the Social Sciences and Humanities. They have governance and support from large governmentally financed organisations and can therefore be seen as trustworthy. Which one/s should be used largely depends on what type of object should be identified and why. It should also be considered if a choice of PID service should be a coherent choice for all members in a (sub-) community, which will possibly limit the choice. Furthermore, it is important to realize that (deep) interoperability between PID systems, above the simple resolution into the object's location, is a difficult matter. For instance making use of the with a PID tightly coupled metadata such as an object checksum, possible with the EPIC services, will be difficult to realize when the underlying PID technology is different. The same holds for switching between service providers when the business plan of one PID service becomes unattractive. In both cases it is desirable that this underlying PID technology used is then the same such as in case of DataCite and EPIC.

The promotional and educational material that will be produced in collaboration with WP7 will cover all relevant aspects to consider when choosing which PID service to use.

APPENDIX

1 Questionnaire for the communities

1. Are there any policies and/or routines for PID services at the community level?

2. Approximate how many/what percentage of the data centers in the community use PID services.

3. Are there any additional requirements on the PID services at the community level?

Terminology

PID services: The PID services we want to analyze in this subtask are those that give additional services, not only register PIDs and resolvers. For example, DataCite manage PIDs, and relate metadata to the PIDs that describe the data.
PID service providers: Here: DataCite, EPIC, PersID (URN:NBN-based services)
Community: (In this task: ERICs) Dariah, Cessda, Clarin for example. The area of interest, research area,
Community center: Research center, data center
Data center: a unit responsible for making data accessible and preserved.

2 Questionnaire for the data centres within the communities

The alternatives listed below the questions are there for guidance, and do not have to be answered one by one.

1. Are there any policies and/or routines for the use of PID services at the data center?

2. Does the data center use PID services?

3. If the data center uses PID services: why does the data center use PID services? What purposes and expectations/requirements does the center have?

- Compulsory
- Identification
- Citation
- Data access - Location of data (using resolvers)
- Search

- Additional services to own, internal PID services
- Other

4. If the data center **does not** use PID services: what is/are the main reason/-s?
   If interested in using PID services; are there any obvious obstacles that prevent the data center from using PID services?
   - Technically complicated/Problems with the technical solutions
   - Metadata issues - e. g. missing required metadata
   - Lack of resources - e.g. PID services currently not prioritized
   - Other

   If **not** interested in using PID services**,** why?
   - Lack of knowledge
   - Cannot see the advantages
   - Fear of external control / lack of robustness / dependence
   - Other

5. If the data center **does not** use PID services: what would it take for the data center to start using PID services?
   Actions taken by the PID service providers based on requirements (specification) from the Communities:
   - Extended Services from the PID service providers - e.g. domain specific metadata specified by the communities/data centers, easier-to-use user interfaces
   - Education and training - for example workshops, tutorials
   - Targeted / direct contact - e.g. introductions, seminars at your data center
   - Other

6. What are the pros and cons with those PID Services that are currently recommended/used by the community/center (experiences on both overall and detailed levels)?

7. Other comments regarding PID services:
   - Additional requirements
   - Other

Terminology

PID services: The PID services we want to analyze in this subtask are those that give additional services, not only register PIDs and resolvers. For example, DataCite manage PIDs, and relate metadata to the PIDs that describe the data.
PID service providers: Here: DataCite, EPIC, PersID (URN:NBN-based services)
Community: (In this task: ERICs) Dariah, Cessda, Clarin for example. The area of interest, research area,
Community center: Research center, data center
Data center: a unit responsible for making data accessible and preserved.

3 Previous works on PIDs

There are numerous projects and institutions that have looked into PIDs. Some, but far from all are referenced in this section.

Athena
The Athena report (www.athenaeurope.org/getFile.php?id=725), published July 2010 gives an extended overview of requirements for persistent identification of objects, collections, and institutions. These requirements have been based on earlier specified requirements created by Digital Preservation Europe (http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf).

Athena is part of the eContentplus programme. Very special in this report is the focus on both physical and digital objects. In the setting of this DASISH task, the requirements for PIDs of digital objects are the most relevant ones.

PersID
The PersID project has delivered five reports (http://www.persid.org/documents.html) in 2011. The scope of the project was broad, dealing with multiple types of organisations serving different communities: cultural heritage organisations, data archives and national and academic libraries. Most of the relevant user requirements can be found in the third project report (Current State of the Art and User Requirements, http://www.persid.org/downloads/finalreports/PersID_Report_Part_3_final.pdf).

APARSEN
Another important project that has delivered substantial information relating to user requirements is APARSEN (http://www.alliancepermanentaccess.org). APARSEN has produced in 2012 the report Persistent Identifiers Interoperability Framework (http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22_1-01-1_9.pdf). This report has a strong focus on user requirements of Persistent Identifiers systems.

Digoiduna
On behalf of the European Commission, the University of Trento has conducted a study (http://www.digoiduna.eu) on identifiers for digital objects and authors.
The report also covers some Researcher Identity solutions.
(http://www.digoiduna.eu/home/DIGOIDUNA_final_report_expert_feedback.pdf?attredirects=0&d=1)

CERL
Consortium of European Research Libraries instigated a Report on Persistent Identifier;, Hans-Werner Hilse and Jochen Kothe, *Implementing Persistent Identifiers: Overview of concepts,*

guidelines and recommendations, which explains the principle of persistent identifiers and helps institutions decide which scheme would best fit their needs. (http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8)

CLARIN

The DASISH partner Common Language Resource and Technology Infrastructure's requirements on PID systems; Persistent and Unique Identifiers (http://www.clarin.eu/sites/default/files/wg2-2-pid-doc-v4.pdf)