How to Deal with Persistent Identifiers?

A Survey on PID usage within CESSDA ERIC

Kerrin Borschewski, Brigitte Hausstein | GESIS – Leibniz Institute for the Social Sciences

Wednesday, 29 May 2019 | IASSIST2019 – Sydney, Australia

cessda.eu @CESSDA_Data

Social Science Data & Persistent Identifiers (PID)

- Data intensive social sciences: easy way to identify & locate

PID
- alphanumeric code pointing to a resource
- unique & persistent
- unambiguous referencing, authentication & validation

- PID = prerequisite for sustainable resource discovery

cessda

Illustration: pixabay

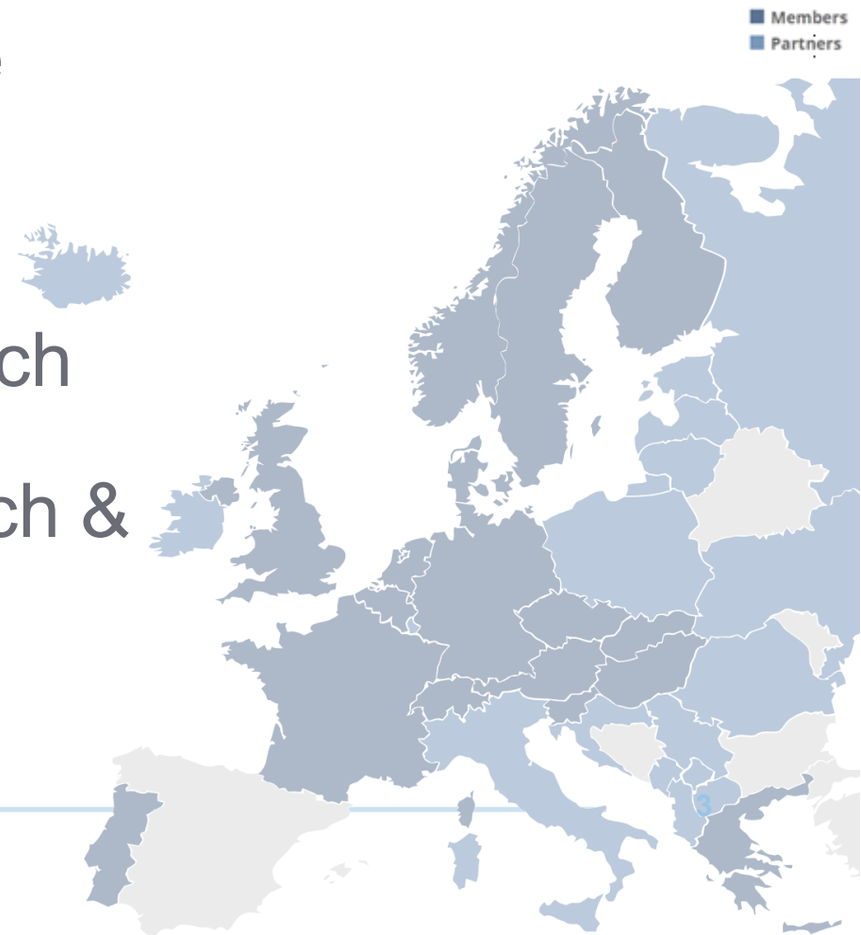
About CESSDA ERIC

About

CESSDA stands for Consortium of European Social Science Data Archives and ERIC stands for European Research Infrastructure Consortium.

- Large-scale, integrated & sustainable data services
- Social science data archives across Europe
- Aims:
  - promoting results of social science research
  - supporting national & international research & cooperation.

https://www.cessda.eu/



CESSDA Online Survey on PID (2015)

Illustration: pixabay

Online Survey Background Information

- Survey period: 8th October – 30th November 2015
- Questionnaire:
  - All CESSDA member organizations
  - Determine status quo for PID usage & demands to a CESSDA PID policy
  - Completed by 11 out of (at that time) 14 CESSDA Service Providers (SP)

Use of PID in CESSDA in 2015

NON-USER (5 CESSDA SP)

FSD
URN:NBN

LiDA
DOI
Handle

NSD
DOI

ADP
URN:SI:UNI-LJ-FDV:ADP

FORS
DOI

USER (6 CESSDA SP)

ČSDA
DOI (2015)
Handle (?)

DDA
DOI (2013)

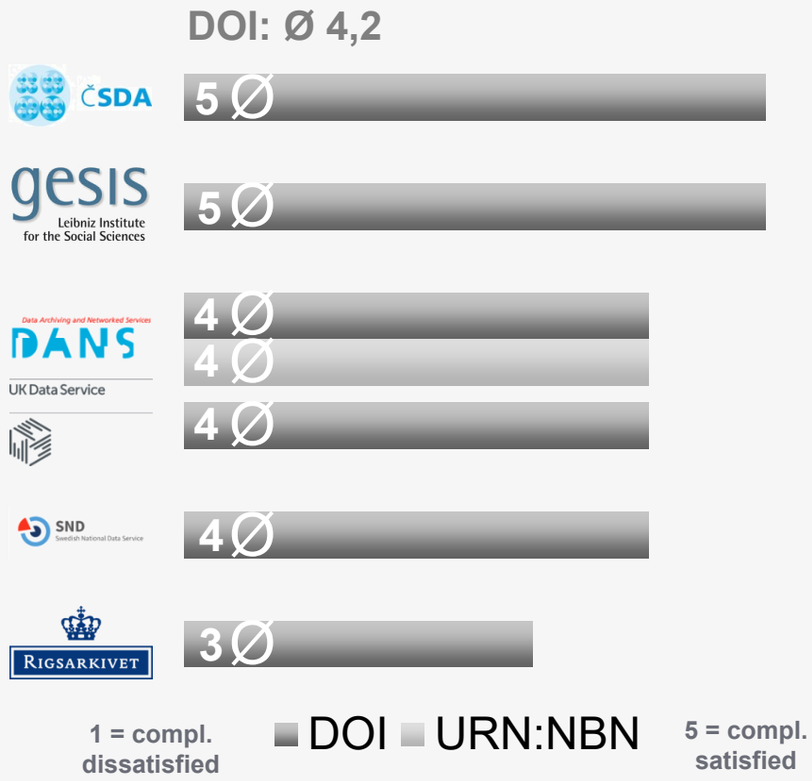
GESIS
DOI (2010)

DANS
URN:NBN (2007)
DOI (2015)

SND
DOI (2012)
ePIC (?)

UKDA
DOI (2011)

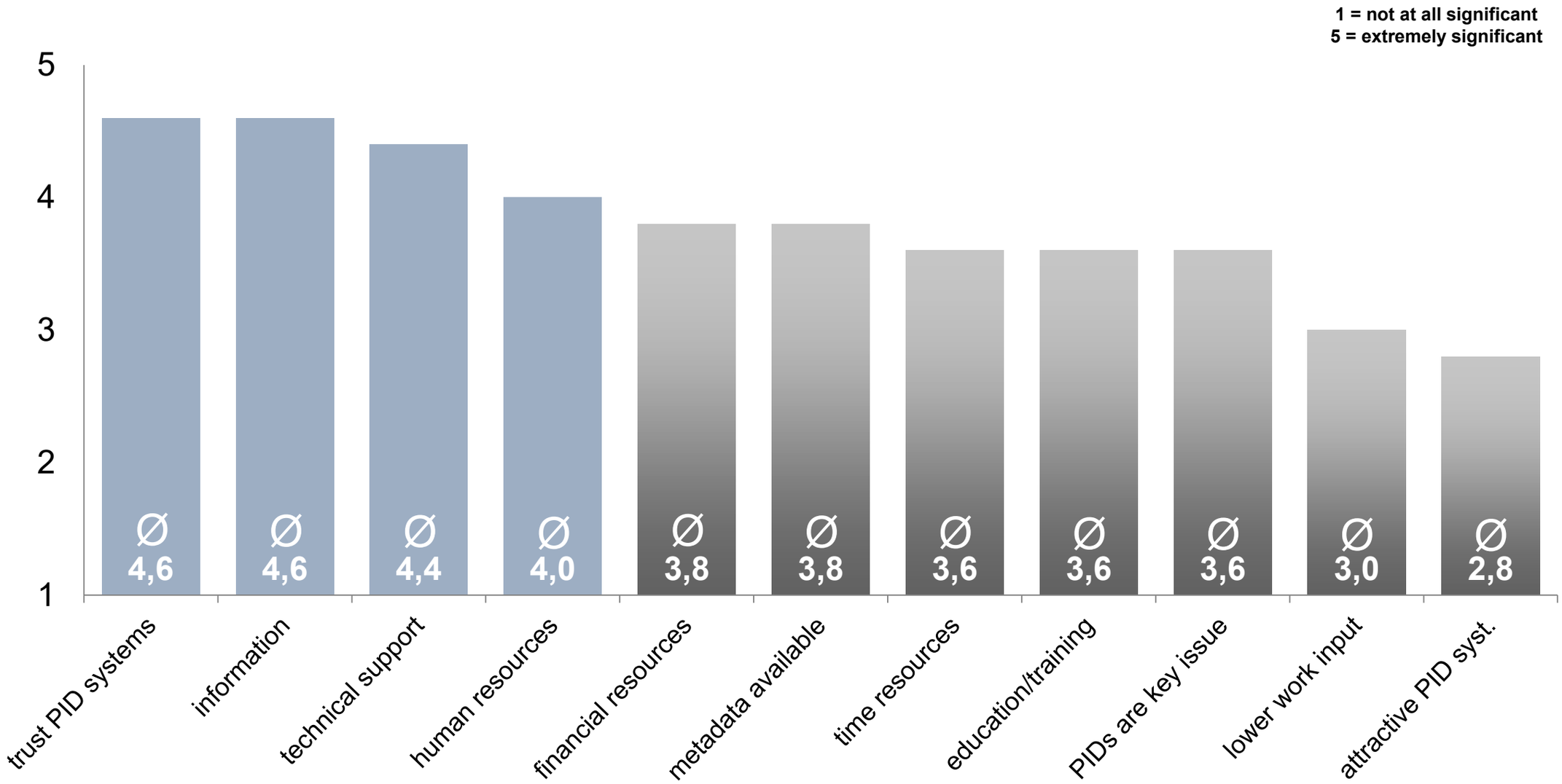
Satisfaction with used PID



Question 2 (User): Please indicate which of the persistent identifier(s) your organization uses (for its data holdings), and how satisfied your organization is with (each of) your used persistent identifier(s)? Answers provided by: ČSDA, DDA, GESIS, DANS, SND, UKDA
Question 4 (Non-User): Which PID system(s) is your organization considering (for its data). Answers provided by: FSD, LiDa, NSD, ADP, FORS

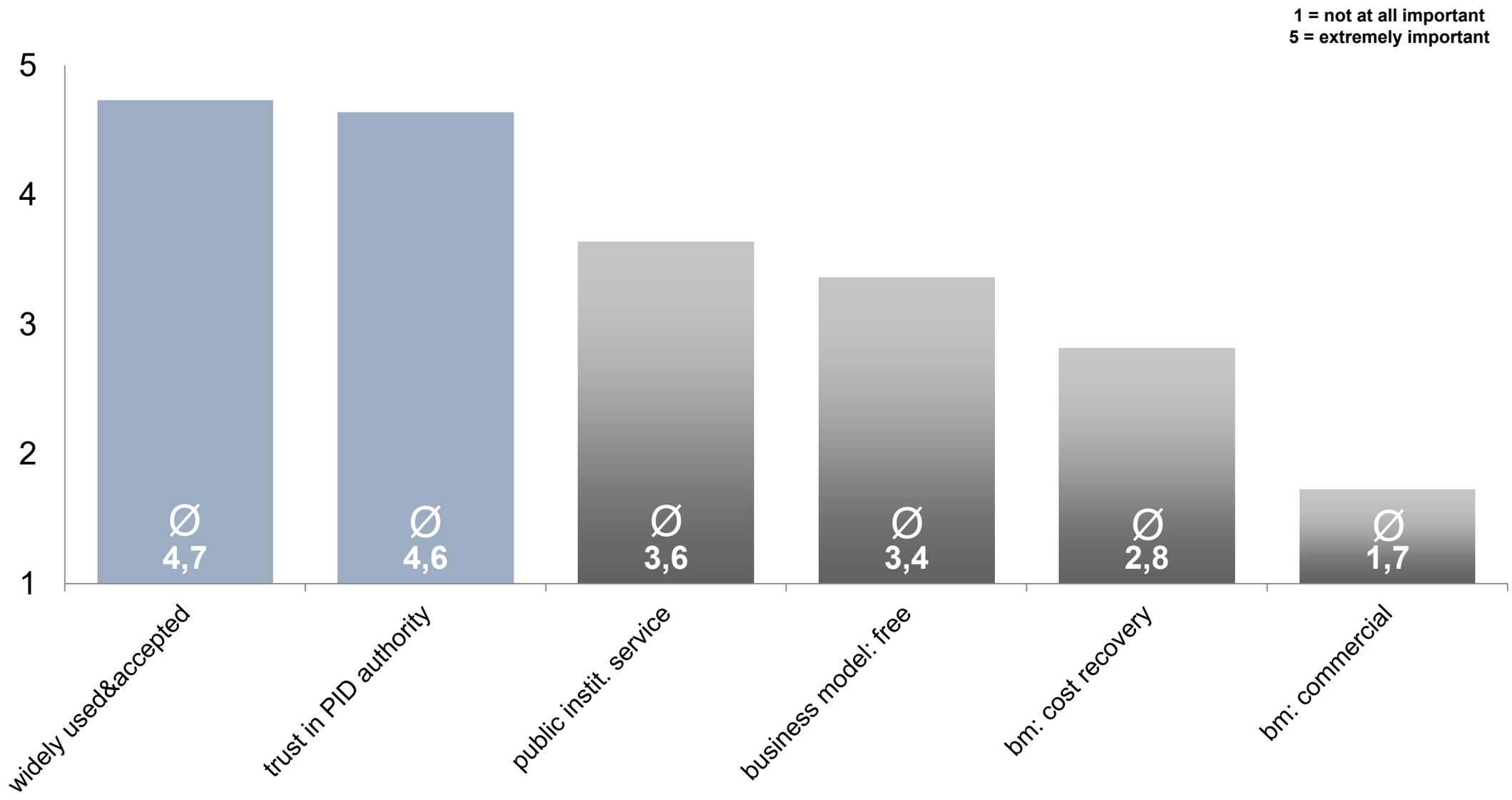
Illustration: pixabay

Non-User's Needs for PID System



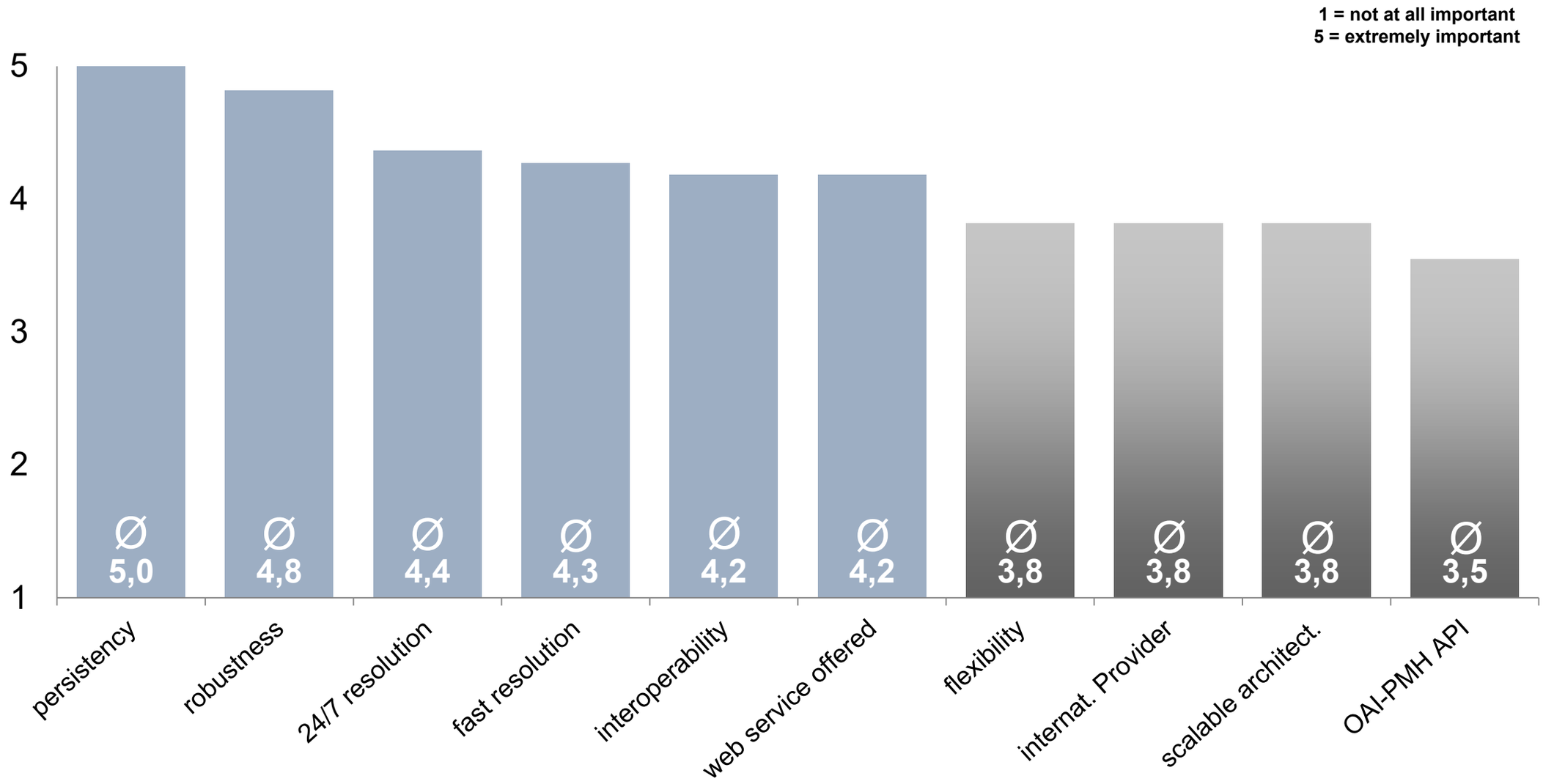
Question 5 (Non-User): Think about your organization: how significant are the following needs when considering introducing a PID system?
Answers provided by: FSD, LiDA, NSD, ADP, FORS

General Expectations “Authority & Credibility” (user + non-user)



Question 6 (ALL): Please indicate how important each of the following authority and credibility issues is for your organization?
Answers provided by: ČSDA, DDA, FSD, GESIS, LiDA, DANS, NSD, ADP, SND, FORS, UKDA

General Expectations “Architecture & Infrastructure” (user + non-user)



Question 6 (ALL): And indicate how important each of the following architecture and infrastructure issues is for your organization, when considering a PID system?
Answers provided by: ČSDA, DDA, FSD, GESIS, LiDA, DANS, NSD, ADP, SND, FORS, UKDA

Open Question on Associated Metadata (user + non-user)

- Dissent: metadata for PID?
- If metadata: different information considered important
  - Range: From minimal metadata information to full citation metadata and more

Question 7: What is the position most prevalent at your organization, which information needs to be covered by metadata that is connected to a PID?
Answers provided by: DANS, GESIS, LiDA, NSD, ADP, SND

Illustration: pixabay

Open Question on CESSDA PID Policy (user + non-user)

- Flexibility – all SP able to implement policy
- Archives to choose their preferred system
- No financial drawbacks
- Interoperability of policy

Question 8: Please note here any further remarks or concerns you have about a common CESSDA PID Policy.
Answers provided by: DANS, FORS, GESIS, NSD, ADP

Illustration: pixabay

CESSDA Expert Interviews on PID (2015)

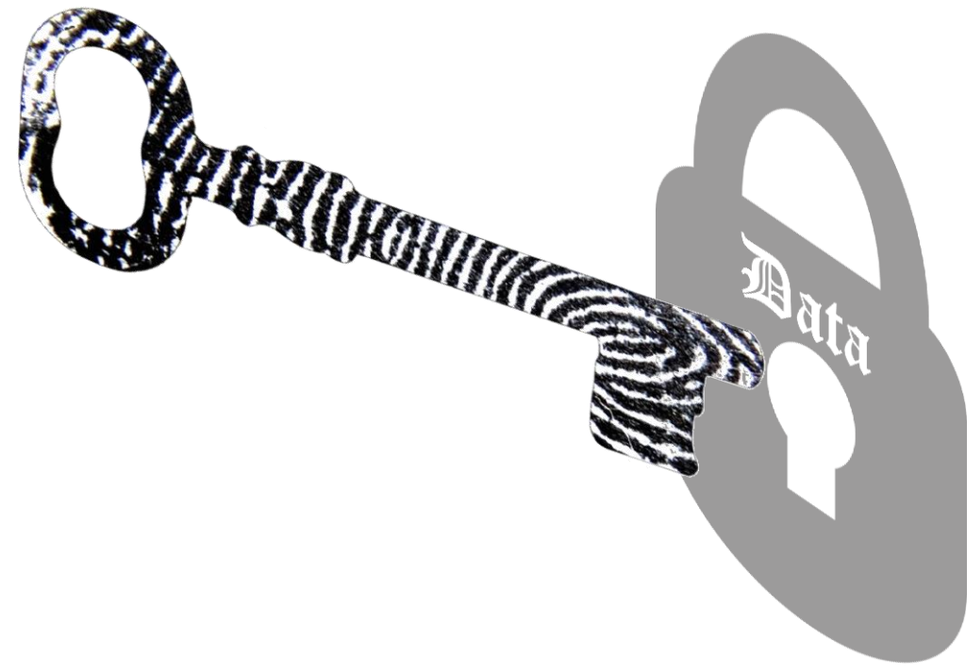


Illustration: pixabay

Expert Interviews Background Information

Survey period: 19th October – 13th November 2015

Experts from:

gesis Leibniz Institute for the Social Sciences

SND Swedish National Data Service

Data Archiving and Networked Services DANS

Topics of main interest:

- Criteria for a common CESSDA PID Policy
- Contents of a common CESSDA PID Policy

Criteria for CESSDA PID Policy

- Short, precise & easy to understand
- Focus on end-user
- Allow different PID systems
- Align with other PID policies
- Take diversity of SP into account
- Guidelines to ensure similar assignment of PID

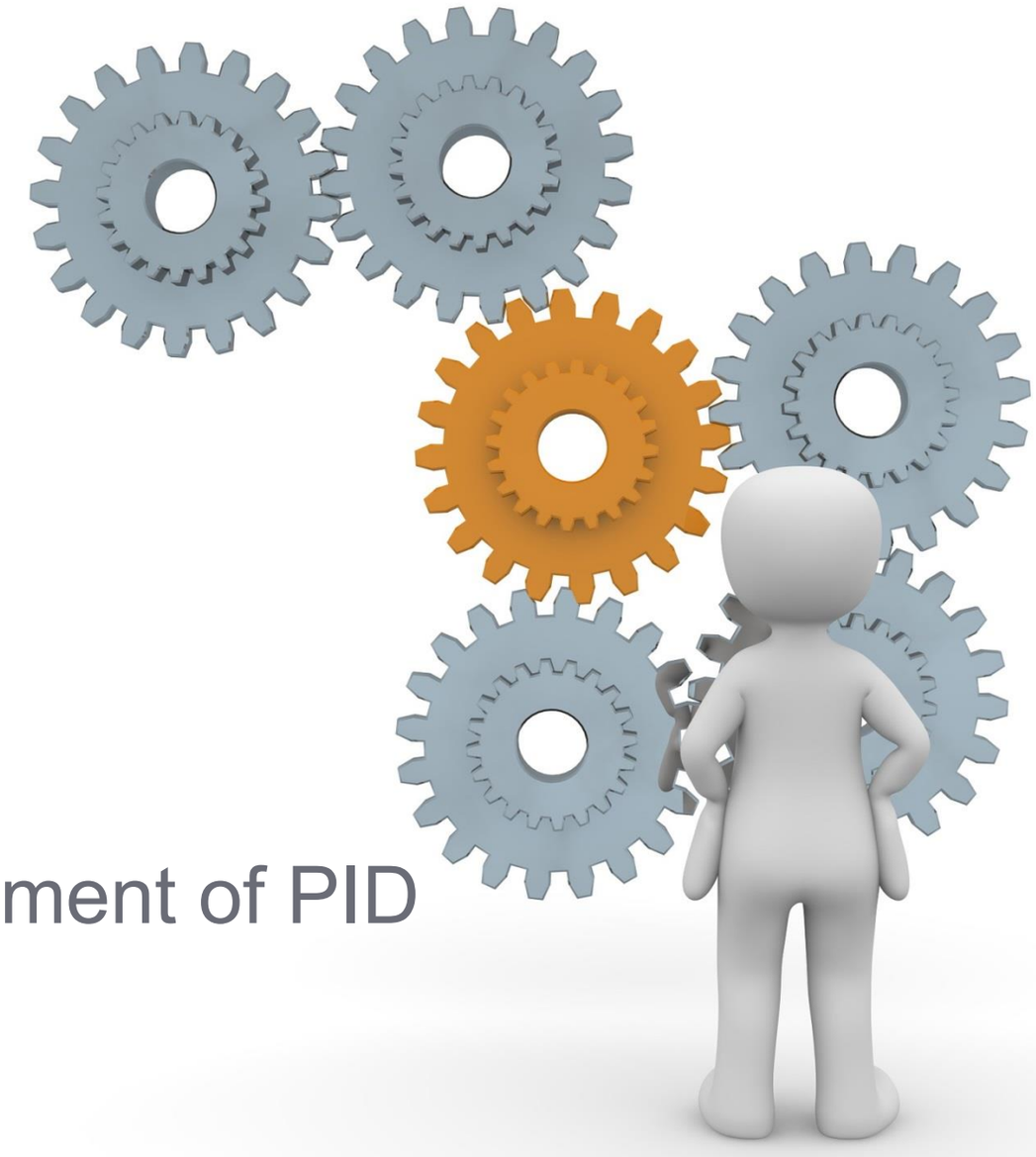
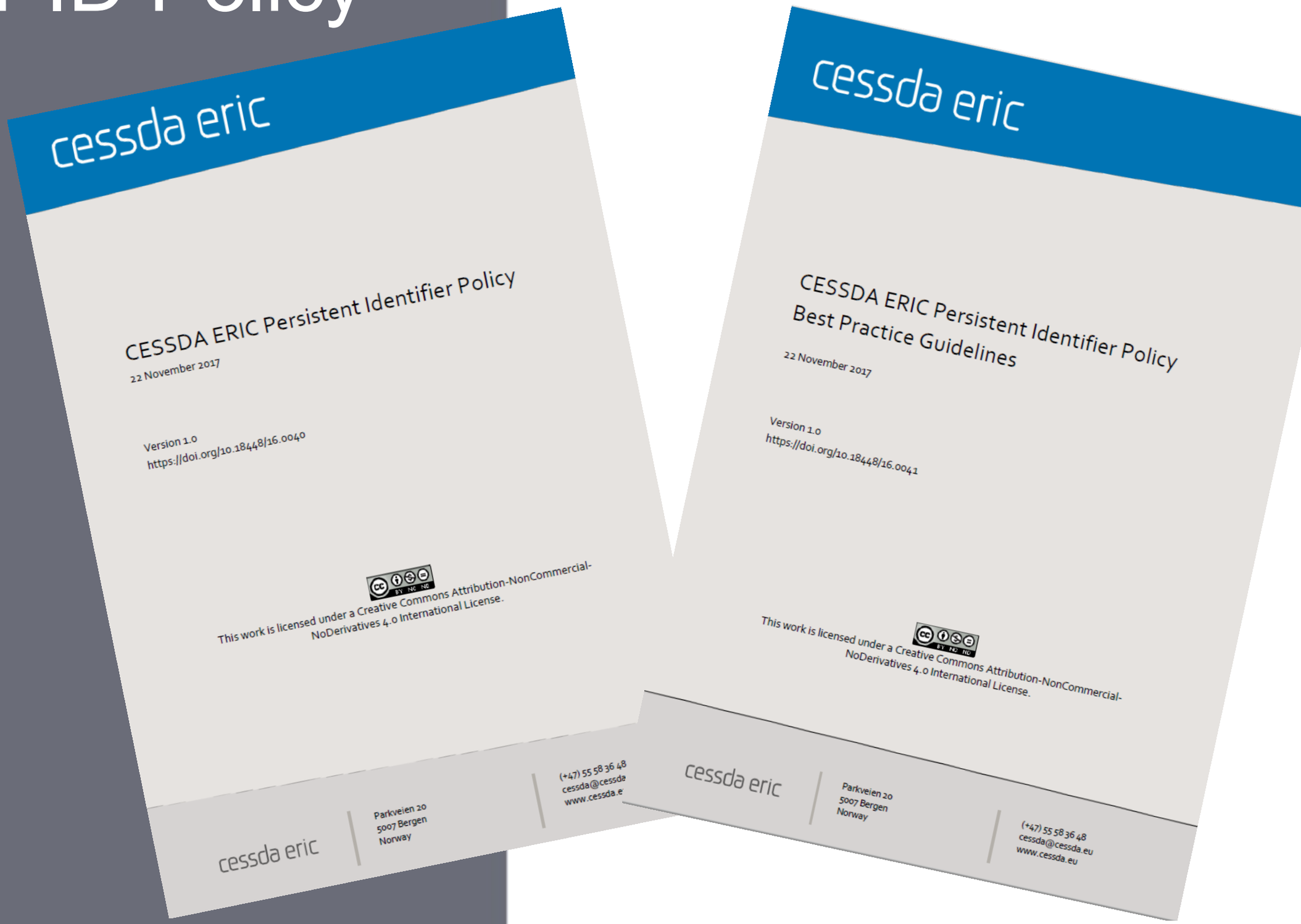


Illustration: pixabay

The CESSDA ERIC PID Policy



Policy: https://doi.org/10.18448/16.0040
Best Practice Guidelines: https://doi.org/10.18448/16.0041

PID are Key – CESSDA PID Policy

cessda eric

CESSDA Persistent Identifier Policy

PRINCIPLES

Principle 1 – Identifying

Each CESSDA SP shall use globally unique and persistent identifiers to identify their data holdings, which are of interest to CESSDA.

Principle 2 – Locating

All data holdings of each CESSDA SP shall be findable by their global PID via the Internet.

Principles 3 – Resolving

CESSDA SP shall use global PID services that ensure 24/7 resolvability of PID.

Principles 4 – Referencing and Citation

PID need to be used to ensure referencing and citation of the data holdings of each CESSDA SP.

Principle 5 – Visibility

PID must be included in the resource-discovery metadata provided by CESSDA SP for the Product and Service Catalogue (PaSC).

Principle 6 – Flexibility

This Policy shall be reviewed at least every two years and adjusted according to the latest strategic and technological developments within and outside of CESSDA.

Policy: https://doi.org/10.18448/16.0040

Illustration: pixabay

CESSDA PID Best Practices Guidelines

cessda eric

1 INTRODUCTION

1.1 Purpose of the document

This document contains guidelines for the implementation of the CESSDA Persistent Identifier Policy Principles. These guidelines address all CESSDA Service Providers (SP). This information is aimed at supporting CESSDA SP to implement the use of global Persistent Identifiers (PID) and to assess the consequences for their organisation.
We have chosen for short descriptions in the document to keep it readable. Suggestions for further reading and specialised documentation are given in the appendix.

1.2. Related Document

CESSDA ERIC Persistent Identifier Policy. Version 1.0. 22 November 2017. https://doi.org/10.18448/16.0040

2 BACKGROUND

As social sciences tend to be more and more data intensive, data repositories must facilitate several ways of identifying and locating data. This development poses complex technical and organisational challenges to data providers. Persistent identification is becoming a prerequisite for sustained and reliable resource discovery and reuse. The use of PID is an important feature of a certified and trustworthy data repository. PID support access to data as well as referencing and citing data. They are an advertisement for data integrity – ultimately PID are part of the proof that an object which a repository has responsibility has not changed. Additionally, the use of PID helps data repositories to be compliant with the FAIR principles (Findable - Accessible - Interoperable - Reusable)[1] set by FORCE 11[2] and provides a future-proof plan in case of the relocation of its holdings.[3]

The main task of CESSDA and its Service Providers is to provide their designated communities with well documented, verifiable, and understandable data for research. One way of doing this is to assign a PID to the data collection (and if desired even to other data collection related objects) and their later versions. The PID will accompany the specific version of the data collection: it is assigned inside the repository and displayed (cited) outside when the data collection is reused. This allows for keeping track of which dataset version is disseminated and gives the user a simple way of both citing the data creator and displaying which exact version was used.

To achieve this goal the assigned PID must be unique on a global scale and the PID service provider must be a trusted organisation with a clear policy on the long-term support of the service and a sustainable business model. The PID assigner (a CESSDA SP) must provide landing pages for each assigned PID with information about how to access the data, licensing rules, different versions and provenance. Some PID services make use of additional metadata alongside their PID which can contain information about related material such as publications or other data collections. If additional metadata is used, this should be made available on the landing page as well.

3 PRINCIPLES

The following pages contain a general description as well as recommendations and examples of the six CESSDA Persistent Identifiers Policy Principles.

[1] FAIR Data Publishing Group: https://www.force11.org/group/fairgroup
[2] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014. https://www.force11.org/group/joint-declaration-data-citation-principles-final
[3] See: Appendix B to this document: CESSDA PID Task Force: PID - What are the benefits for CESSDA community (users and providers of services)? Cologne, 26 September 2016.

cessda eric

Principle 1: Identifying

Each CESSDA SP shall use globally unique and persistent identifiers to identify their data holdings, which are of interest to CESSDA.

General information

A globally unique and persistent PID is a worldwide unique and persistent archival reference code. It is an identifier that is permanently assigned to an object and unique on a global scale. Persistence is a commitment by the issuing PID service provider ensuring that the system is administered comprehensively.
Attaching a unique identifier to a data collection helps the data provider to guarantee the origin, content and version of the data collection. This can be verified by resolving the PID. In order to work on a global scale the service must run 24/7 and must be supported and governed by a dedicated user community (see also principle 3).
The main PID systems for data collections are: Handle, DOI®, URN:NBN and ARK.
DOI® and Handle are based on the same (Handle) system; URN:NBN on a network of services from National Libraries and ARK is a service developed by the California Digital Library (CDL).
DOI® and Handle are the ones most commonly used.
In terms of identifying, by using PID, CESSDA SP will be able to

- identify their holdings persistently on the Internet so that these holdings can be referenced
- keep track of disseminated data and their versions
- provide a simple way of citing and referencing both data collection and data creator
- add a building block towards the status of a Trusted Digital Repository[4].
- be compliant with the CESSDA Data Access Policy[5].

Recommendations

- CESSDA SP should assign globally unique PID to data that are published.
- CESSDA SP should commit to maintaining the PID permanently.[6]
- The assigned PID must always resolve to the same object.
- The PID should preferably be embedded within the data file (e.g. as a variable).
- Which PID service a CESSDA SP may use depends on various reasons (e.g. harmonisation with other national or already existing services etc.).
- When choosing a PID system and a PID service provider[7], credibility and long-term viability of the PID system as well as the use of open standards should be considered. CESSDA SP should use PID service providers with clear and transparent policy and business models.

Examples

PID	Example PID name
Handle	hdl://hdl.handle.net/11022/0000-0000-0000-C
DOI®	https://doi.org/10.5878/002645
URN:NBN	urn:nbn:de:bvb:19-146642
ARK	http://bnf.fr/ark:/13030/tf5p30086k

[4] For more information on Trusted Digital Repositories see: Edmunds, R; L'Hours, H.; Rickards, L.; Trilsbeek, P. and Vardigan, M. (2016): Core Trustworthy Data Repositories Requirements. https://doi.org/10.5281/zenodo.168411
[5] http://cessda.net/eng/About-us/Documents
[6] For further information on the provision of tombstone pages see Principle 5
[7] See: Appendix B: CESSDA PID Task FORCE: Review of the PID Services provided by GESIS, DANS and SND. Cologne 29 August 2016.

Illustration: pixabay

Outlook

- Revision & update of PID Policy and Best Practice Guidelines
- Assist CESSDA ERIC SP with PID issues
- Involvement in developments/initiatives around the globe
- Granularity issue of PID assignment

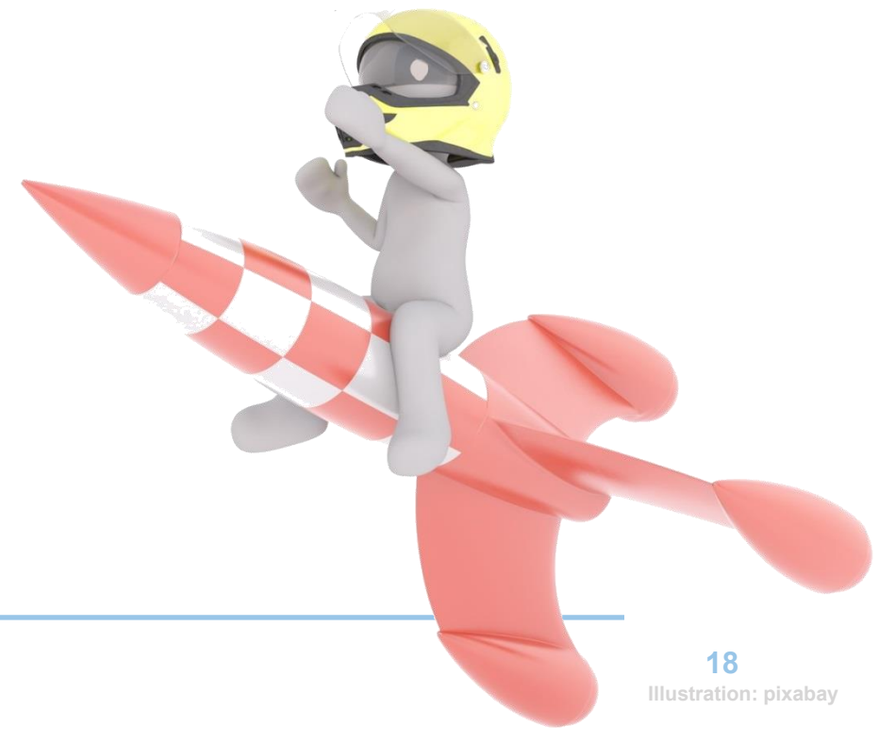


Illustration: pixabay

Stay curious and enjoy Australia

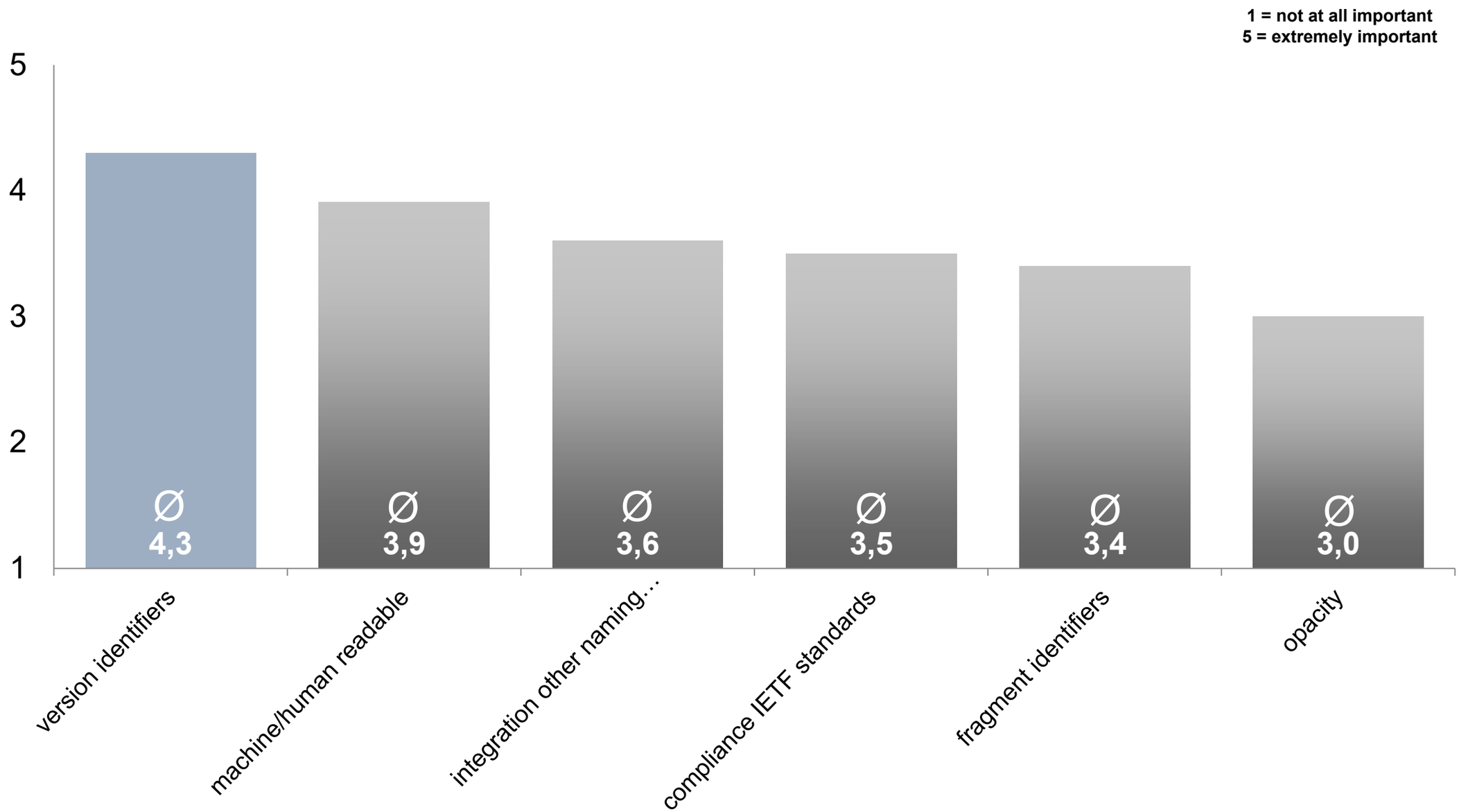
cessda

Image: pixabay

Appendix

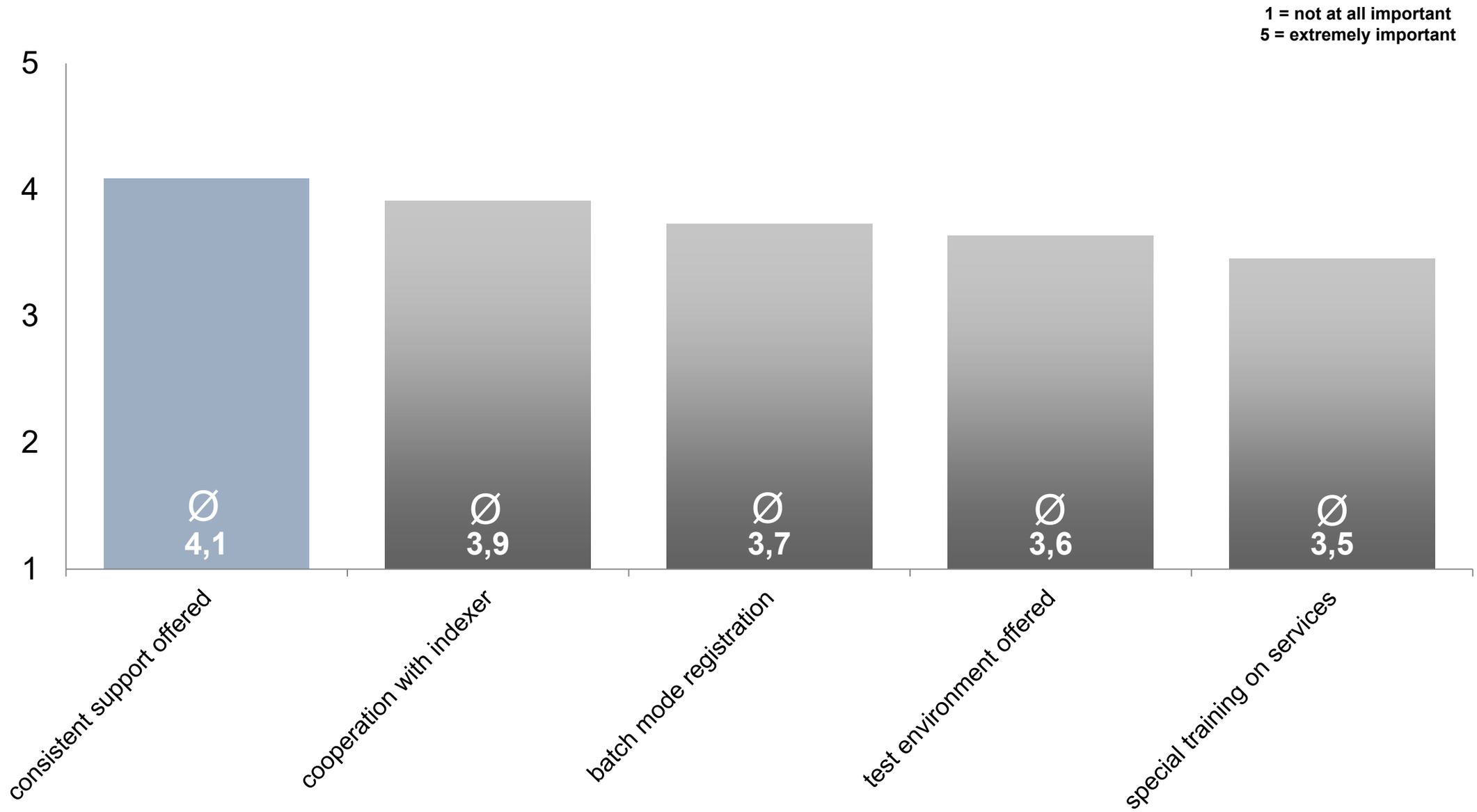
Image: pixabay

General Expectations “PID Syntax” (user + non-user)



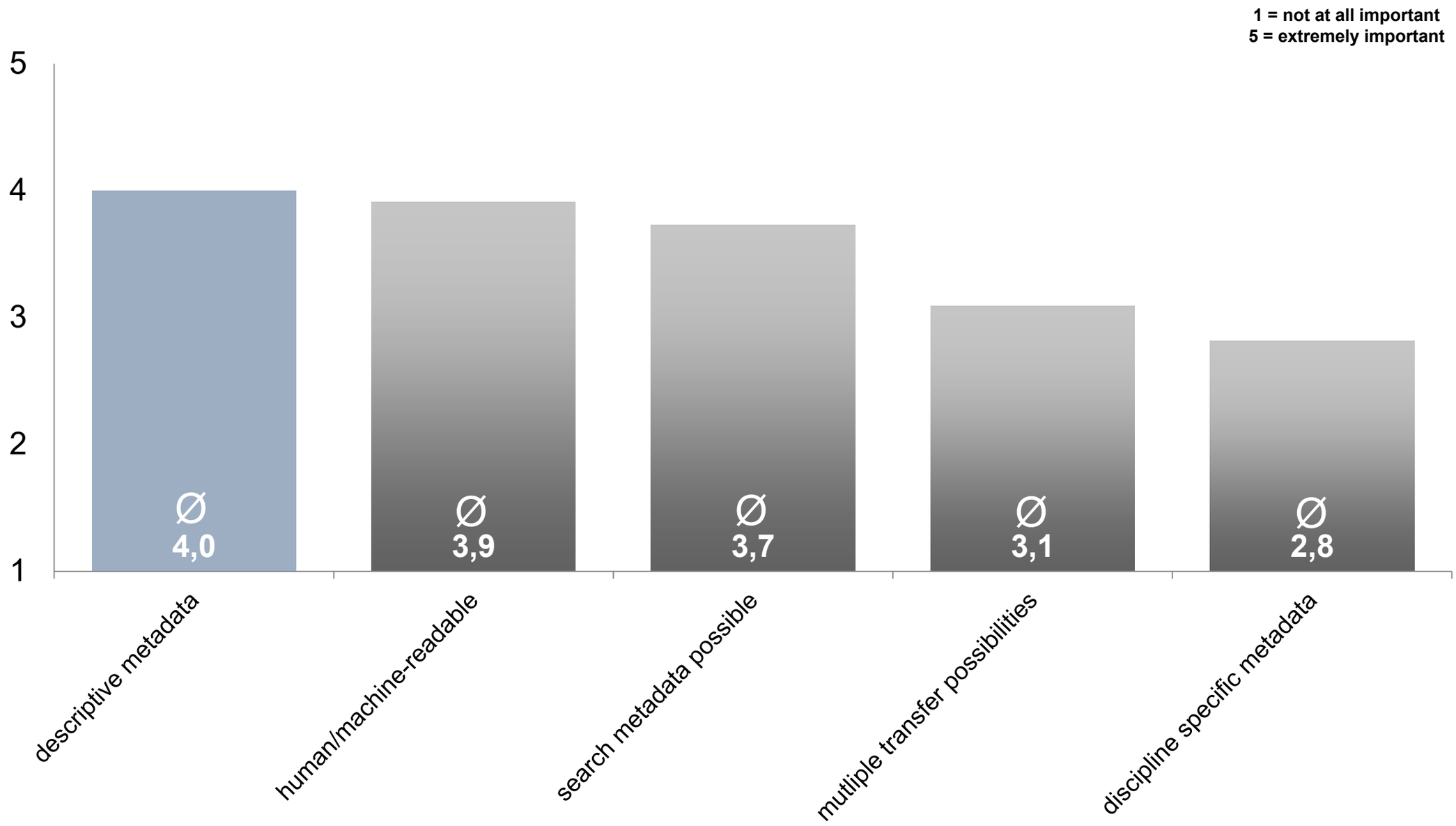
Question 6 (ALL): And indicate how important each of the following PID syntax issues is for your organization, when considering a PID system?
Answers provided by: ČSDA, DDA, FSD, GESIS, LiDA, DANS, NSD, ADP, SND, FORS, UKDA

General Expectations “Additional Technologies” (user + non-user)



Question 6 (ALL): And indicate how important each of the following technological aspects is for your organization, when considering a PID system?
Answers provided by: ČSDA, DDA, FSD, GESIS, LiDA, DANS, NSD, ADP, SND, FORS, UKDA

General Expectations “Metadata” (user + non-user)



Question 6 (ALL): And indicate how important each of the following metadata issues is for your organization, when considering a PID system?
Answers provided by: ČSDA, DDA, FSD, GESIS, LiDA, DANS, NSD, ADP, SND, FORS, UKDA