Towards a Global Digital Object Cloud – Report from the Views on PID Systems training course and workshop

Editors: Kathrin Beck, Raphael Ritz, Peter Wittenburg

The training course and interaction workshop organised by RDA Europe from 31.8 until 2.9.2016 in Garching/Munich brought together 47 experts from different organisations and initiatives from 13 countries[1] [1]. All those parties were dealing in some form with the usage of persistent identifiers (PIDs) in practical applications – be it for referencing or citation purposes. In addition to experts from initiatives that are known for their key role in this field, participants came, e.g., from infrastructure builders, domain repositories and library environments. To link up with industry, the chief of Operations and Planning Department of International Telecommunication Union (ITU) [2, 3] was invited as well. The three intensive days were meant a) to update data professionals on current principles and usages and b) to interact on applications of PIDs and the necessary development of PID systems in the future. This report presents the core messages from these three days as agreed by the participants focusing on data management, access and reuse scenarios.

PID Concept

In a phase where data intensive science is moving towards automated processes, the usage of PIDs for any type of Digital Object [4] is crucial. This basic message has been accepted throughout the scientific disciplines and is also of key relevance across other sectors such as different industries, Internet of Things (IoT), governments, health, etc. and one can observe how this concept is taken up broadly and turned into useful services such as provenance tracking of medical trial data, cross-departmental services in universities, supporting the powerful notion of collections, and many more. The need for PIDs is hardly questioned anymore, but now it is time to share knowledge, experiences and services, and make our PID landscape more mature and interoperable.

Proper PID usage and support will become key for competitiveness in science and industry.

New Opportunities

After 20 years of experience, it has been shown convincingly that the systematic usage of PIDs opens up completely new possibilities to develop added-value services, be it in reliably linking all kinds of information and data, in exactly tracking digital objects or digital representations of physical objects in process chains, or in checking of identity and integrity of digital objects after long-term storage, etc. Though, offering new opportunities, the area of PIDs is complex and amongst scientists as well as in industry, there is still much confusion about PIDs and PID systems requiring a phase of simplification by clear and simple messages.

PIDs need to be used by all parties dealing with data professionally to make full use of advanced opportunities. A PID centric approach to data management, access and use will open the way towards new and comprehensive way of data handling and finally to a Global Digital Object Cloud [5] as a generic, non-proprietary virtualisation layer.

National Task

20 years ago PIDs were meant to be associated with a modest number of digital publications. However currently industrial and scientific applications are already creating millions of PIDs within days. We are now working towards a global infrastructure for all kinds of applications in research, industry and society requiring structuring and maturing the PID service landscape. As in the phase

[1] All slides can be found on the training/workshop web site.

when the internet enabled national services, we need to work towards well supported and structured global and national PID services with high availability and capacity that allow people to rely on them and that will be the basis for a PID centric approach to data management, access and use.

International and national steps need to be taken urgently to offer a sustainable, structured and mature PID service landscape based on quality assessed service providers to all interested parties. Only such a structured and massive approach will prevent ending up with unresolvable PID zombies.

Solution Space

The number of available solutions to build the basis for a Global Digital Object Cloud has been drastically reduced to the space offered by the Handle System. Trustworthy service providers such as those offering DOIs (Digital Object Identifier) [6] or EPIC Handles [7] organise communities of use and the underlying Handle System [8] is now governed by the independent DONA Foundation [9]. Efforts need to be taken to make the services based on Handles available across services providers. Also in industry the Handle System as basis of the digital object architecture gets attraction.

PIDs are becoming essential across sectors and communities for different application scenarios and efforts need to be taken to offer services across these sectors and communities.

Role of Repositories

Trustworthy repositories that comply with quality criteria are at the source of a functioning data landscape. They need to assign stable PIDs and metadata to all uploaded digital objects and are responsible for maintaining the PID records and the relationships amongst the key elements of a Digital Object (PID, metadata, locations of bit sequences, landing pages, etc.). In particular, the relationship between PIDs and high-quality metadata is needed to allow both machines and humans to make proper interpretation of the identified objects. In order to ensure that the PIDs repositories will not become zombies, repositories should focus on training, communication on PID , new services around PID and particularly on listening to the requirements and needs of end users.

Setting up and maintaining trustworthy repositories is key for a structured data landscape guaranteeing access to data and its accompanying metadata.

Automation

New developments such as in IoT for example will increase the amounts of data and its inherent complexity to a scale that only automatic procedures will allow us to make use of. As early as possibly we need to assign PIDs for all entities involved, which will be key to success. In future all kinds of data from all domains, like environment, climate and health, will be connected and used for analysis.

We need to design the required mechanisms and build the needed tools now with high urgency.

DOI/Handle Service Providers

Until now we have seen that, due to existing needs, many Handle-based services providers have been established servicing different needs resulting in the need for quality control and the addressing of cost issues. It is time to structure this landscape by introducing a Quality Seal and to make clear what can be expected from the various Handle services.

We urgently need to come to a structured and integrated domain of Handle Service Providers.

Referencing and Citing[2]

The discussion identified two major use cases: (1) Stable and persistent referencing of digital objects is a must for proper data management and access strategies and labs and industry are increasingly

[2] The term „referencing“ is meant here in the computer science sense as an actionable pointer that can be used by machines.

often use Handles from trustworthy service providers. (2) For the citation of digital objects as part of the scholarly record most often DOIs are being used.
Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services.

Messages to Users

PID service providers need to turn to simple and clear messages to users that are not driven by competition, but by the insights that we need an integrated worldwide PID domain based on Handles which everyone can use and rely on, just in the same way as we have an integrated worldwide IP domain to connect network devices. It will be the case that for special applications other systems will be in place and that we will have to take care of legacy solutions.
The PID centric approaches that are key to manage the data Tsunami require simple and clear messages for the users.

RDA's Role

Through its inclusive and global approach, Research Data Alliance [10] has shown its capability to bring people together from different communities and now also from different sectors to develop key pillars of the emerging data landscape. Individual results of working groups and activities of groups that combine these results are at the base of what is being stated in this paper.
RDA needs to continue to play its integrative role in the data domain.

References

[1] PID Workshop Pages: https://rd-alliance.org/views-about-pid-systems-training-course-and-workshop-31-august-2-september-2016-garchingmunich
[2] ITU, Platform for interoperability of heterogeneous identity management systems, IOT2020, https://rd-alliance.org/groups/gede-group-european-data-experts-rda
[3] ITU, Relationships between Recommendations ITU-TX.1255(09/2013), ITU-TX.660(07/2011) and ITU-TX.672(08/2010), https://rd-alliance.org/groups/gede-group-european-data-experts-rda
[4] DFT Core Terms and Model: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[5] Global Digital Object Cloud: https://rd-alliance.org/group/data-fabric-ig/wiki/global-digital-object-cloud
[6] DOI: https://www.doi.org/
[7] EPIC: http://www.pidconsortium.eu/
[8] Handle System: https://en.wikipedia.org/wiki/Handle_System
[9] DONA Foundation: https://www.dona.net/
[10] Research Data Alliance: http://rd.alliance.org