**GEDE PID Focus Area meeting**

1 February 2017

12.00-14.00 CET

**Attending:** Peter Wittenburg, Chair (RDA/MPI), Rebecca Grant, Minute Taker (RDA/National Library of Ireland), Antonio Rosato (INSTRUCT), Brigitte Hausstein (CESSDA), Carlo Maria Zwölf (VAMDC), Christine Staiger (SURFsara), Danielle Couvreur (MYRRHA), Damien Boulanger (IAGOS), Dieter Van Uytvanck (CLARIN), Dimitris Koureas (DiSCCo), Joe Padfield (NGL), Margareta Hellström (ICOS), Øystein Godøy (SIOS), Tobias Weigel (DKRZ), Wolfgang Kuchinke (ECRIN)

**Apologies:**

Tamas Gaizer (ELI), Rob Hooft (ELIXIR), Jan-Willem Boiten (EATRIS) and Ingemar Häggström (EISCAT)

**Introduction:** Peter noted that RDA Europe participated in a light review with the commission before Christmas, and that the RDA Europe team have been working to prepare a document in response to the review. Peter will circulate this document to GEDE for information. A document has also been prepared in response to Peter’s meeting with Robert Jan Smits (Director-General for Research and Innovation, European Commission). Peter will circulate this and request comments, feedback and support.

The purpose of the PID Focus Area work is

1. to identify whether there are any other PID-related assertions that should be included in our list; to discuss and understand the assertions;
2. and to find agreements and disagreements regarding the assertions.

The intention of the meeting is to address the first two aspects of this task.

It was also clarified that the contributors to the PID Focus Area do not need to be members of the main GEDE group. Rebecca should be informed about the names and e-mails of all who wish to participate, so that she can update the mailing list.

**Sources of PID assertions:**

Additional assertions were suggested by the group during the meeting. Peter requested that the proposers write a short summary of each of these assertions to be incorporated in the list.

People suggested a list of work related to PID and data services existing in some data communities:

* Christine: “10 Simple rules for design, provision, and reuse of persistent identifiers for life science data”, see <https://zenodo.org/record/18003#.WJHDt7YrJbU>, “Introduction to Persistent Identifiers” by the Netherlands Coalition for Digital Preservation (NCDD), see <http://www.ncdd.nl/en/pid/>
* Dimitris: “Biodiversity Information Standards (TDWG)”, see <https://github.com/tdwg>, &the TDWG Globally Unigue IDentifiers (GUID) applicability statements page, see <https://github.com/tdwg/guid-as>
* Joe: the International Council of Museums Committee for Conservation (ICOM-CC) website <http://www.icom-cc.org/> (noted that PID information could not be located on this website); the International Committee for Documentation (CIDOC) Reference Model, see <http://www.cidoc-crm.org/> (noted that this website does not appear to be secure), and “Statement of principles of museum documentation” from ICOM’s International Committee for Education and Cultural Action, see <http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/DocStandards/Statement6v2EN.pdf>
* Dieter: CLARIN PID policy summary, see <https://www.clarin.eu/node/3965>, and “CLARIN DataCite workshop conclusions”, see <https://www.clarin.eu/content/datacite-workshop-conclusions>
* Wolfgang: CORBEL (COordinated Research infrastructures Building Enduring Life-science services) statements (not public yet, but should soon be available at <http://www.corbel-project.eu/>)
* Carlo: The International Virtual Observatory Alliance (IVOA) document “Table Access Protocol”, see <http://www.ivoa.net/documents/TAP/>

The following resources were also mentioned:

* Report from Dynamic Data Meeting (2014): “Data citation and digital identifiers for time series data / environmental research infrastructures”, see <https://www.bodc.ac.uk/about/outputs/presentations_and_papers/documents/datacitation_juck.pdf>
* The International Telecommunication Union (ITU) recommendation “Framework for discovery of identity management information” X.1255: <https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-X.1255-201309-I!!PDF-E&type=items>
* RDA Europe “Views on PID Systems” PID Training Course (<https://www.rd-alliance.org/views-about-pid-systems-training-course>) and Workshop (<https://www.rd-alliance.org/views-about-pid-systems-workshop>), held in Garching, Germany 30 August - 1 September 2016.
* ”How dead is the PID Zombie zoo?” by R. Huber: <https://www.rd-alliance.org/sites/default/files/attachment/20160902-RDA_EU_View_on_PID_Systems_Garching-Robert_Huber-Jens_Klump-How_dead_is_dead_in_the_PID_Zombie_zoo.pdf>
* PID Centric operation: <https://www.rd-alliance.org/group/data-fabric-ig/wiki/df-configuration-pid-centric-data-management-and-access.html>

**Discussion of PID assertions:**

The assertions were discussed simultaneously on the conference call and via the chat function of the meeting software. The chat discussions are listed separately below.

There was also a question raised regarding the possibility that the wordings of the assertions can be changed by the group – it was noted that comments, suggestions and suggested replacements could be provided by us, but that the original assertions will not be changed.

The assertions are clustered by topic, and the following comments were captured in relation to the discussion of each.

**Nature of PIDs and PID Systems**

* The was broad agreement on the first three assertions, PID1. RDA DFT1.1, PID2. RDA DFT1.2 and PID3.RDA DFT1.3, however it was noted that in life sciences there is still much confusion about PIDs and disputes about how the PID and Linked Open Data world can be brought together.
* In many disciplines it is cheaper to prepare new data, rather than storing (and assigning identifiers to) existing data.
* It was agreed to make this a discussion topic for the session at the RDA Plenary in Barcelona in April. Interested group members should prepare information to present at this session.
* The stability of PID systems is also considered to be an issue as there is always a risk for PID Zombies, it is the emphasis everyone puts into a global service which will make it sustainable.
* Besides Handles, there are many other PID suggestions such as UUID; to make these resolvable one could use a UUID as a suffix in a Handle and the prefix would identify the service.
* Criteria for trustworthy PID systems are necessary, the question is whether the criteria are meant for the whole eco system of Handle services or for the local systems; in case of the latter some criteria don't apply etc.
* The statement about being "non-profit" is too narrow and does not explain what is really meant - this needs to be re-formulated.
* Making a PID system dependent on the will of a company would be dangerous.
* It is important to have an exit strategy at all levels.
* The global Handle System exists already now on 6 root nodes that share root PIDs, i.e. if one node does not work anymore there would be no problem; so obviously exit strategies need to be available for local systems.
* The ITU recommendation X.1255 is of relevance and should be communicated (see links)

**Relevance of PIDs and PID Systems**

* No significant disagreements were noted.

**Assigning PIDs**

* The RDA Data Fabric Terminology group and the FAIR assertions are expressing the same core messages
* The term "eternal" is seen critical since Data Objects in the Life Sciences are dynamic, and the term “dynamic” and its meaning were discussed by the group. It was also noted that in biology experiments are often repeated as it is cheaper than storing the data, but this can create issues for reproducibility.
* Noted that in biological research, people work with the textual data - and the original imaging data is not used anymore; already the reduction can be done using different algorithms, parameters, etc.
* Noted that “old” observations or datasets should be kept and be accessible in most communities and thus need stable references.
* Granularity remains an issue and use-cases need to be given to make assertion PID12. RDA understandable.
* A high granularity cannot always be recommended. It was noted that one can also create subsets of data at a later stage, but this involves additional effort. Maggie noted that significant effort is only required if separate objects (containing subsets) are created, at which point metadata records have to be provided, a PID minted, and the file curated & stored in a trusted repository etc. In case the subset is creatable "on the fly", and e.g. the associated query is assigned a PID (following the RDA Dynamic Data Citation WG recommendation), there is less effort involved.
* The group discussed the use of fragments and fragment identifiers - they are not really part of the identification (the checksum will only be calculated on the whole DO), however they can be used and addressing (resolving) needs to be done locally by services
* The group should look into the meetings about dynamic data and should include the assertions from Data Citation WG
* We should distinguish the usage of PIDs for citation purposes and the use in workflows (pipelines); the latter require a high granularity.
* The difference between referencing and citing was also discussed at the Munich PID workshop. There will be a session at RDA Plenary 9 in Barcelona where this will be discussed in more detail. GEDE members are encouraged to attend.
* In the case that PIDs are used in pipelines or workflows, one could in principle replace function names. It may be interesting to work with people from RDA reproducibility IG.

**Using PIDs**

* "Retrievable" in assertion PID6.FAIR-A1 means "accessible" and not "searchable" - this often creates confusion
* The whole issue of deletion (of data objects and/or identifiers!) requires further discussion.

**Meeting close:**

The next teleconference will take place in early March, and Rebecca will send a Doodle poll to organise this.

It was also noted that GEDE members are welcome to attend the RDA Plenary 9 in Barcelona on April 5-7th (https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona)

**Chat log with responses from Peter**

**12:30 Joe Padfield (NGL):** *Have people discussed PIDs for concepts that are shared accross multiple fields, such as dates and locations ?*

**12:32 Christine Staiger:** *I am also very much interested in the topic: PIDs for code and compute pipelines.*

**Peter:** Many institutions have set up for example their own Handle Service for several reasons. One important reason is for example producitivity. The climate modeling community for eamplee is creating a huge amount of DOs in one simulation. They associate Handles with all objects so that people can refer to it from workflows etc. and thus need to have a powerful system that allows them to create the amount of Handles efficiently. Since the prefiy is registered at the global root system which is now also distributed, all PIDs can be resolved - so indeed it is distributed at mulitple levels. It is up to the local policy whether you want to mirror your local Handles with another service. The question is of course: what happens if you do not mirror, you may want to have local backup, but in case of high availability as requirement by many researchers, you would get acceptance problems. So participating in a redundant system such as maintained by ePIC may make sense.

**12:40 Joe Padfield (NGL)>** *GIven the development of technology how long does something need to last for us to consider it to be persistant?*

**12:41 Joe Padfield (NGL)>** *This is quite important in any trust issues, organisations would need to be stable enough for the time length agreed*

**Peter:** There is no proper answer to this question I assume and at a certain moment data archeologists will need to do some expensive work. When we discussed these kind of issues in the Max Planck Society, the president agree to give a guarantee that MPS will take care that things exist for 50 years and that if a curcial service will be stopped the MPS would take care that data (PID) would be exported to another reliable service. 50 years is already a good perspective or?

**12:42 Maggie Hellström:** *IMHO, the PID "system" should not be locked to any specific technologies (e.g. back-end databases).*

**Peter:** Indeed we need to distinguish at least 2 levels:

* we have the specifications of a standard such as Handle syntax and procedures
* then you have some technologies that implement the standard and allow us to work

In fact everyone could build another resolver for Handles for example if the speciffications are followed. However, let's be very frank: it needs quite some careful design and testing to be able to deploy a system whih we all are going to rely on. But the principle is correct.

**12:42 Joe Padfield (NGL):** *Should this discussion examine the notion of PID mapping services - how would one breath new live into dieing PID systems*

**12:43 Maggie Hellström:** *@Joe: then we will never have time for anything else! :-)*

**12:46 Joe Padfield (NGL):** *Would it be worth developing simple open source PID systems that can be used by people locally, but are designed to allow easy aggregation to hub services. People want to use things that are easy, so they end up building things locally that then increases risk and decreases interoprability*

**Peter:** This is an important issue for legacy. It is known that there is no general URN resolver (as far as I know) and one could simply integrate all URNs as suffixes into the Handle System which would give it a resolver and persistency. But for the future we need to think of a reliable worldwide functioning system and it is there - we just need to give it the value as we take care that our Internet is functioning. As already indicated it does not make sense to me to develop yet another system. You can register a Handle prefix and setup a small server if you want. Otherwise we will all be busy to rescue some PIDs registered somewhere.

**12:46 Joe Padfield (NGL):** *WHatever model the organisation has, it would need to have a notion of guarenteed service*

**12:48 Joe Padfield (NGL):** *If it is profit or non-profit does not really matter, its the long term stablity of service rather than the company. The system would need to be non-profit but it could be supported by a for profit, but the process for replacing the for profit or non-profit company would nee to be hard wired into the orginal plan*

**12:49 Joe Padfield (NGL):** *The difference is between ownerships and service provision*

**Peter:** Right - we need a guaranteed service since we are dependent on its existence already. To me it is obvious that science alone will not be sufficient to guarantee persistency, we need to get industry on board as user as in the case of Internet. It is the sheer mass of investments that will guarantee survival whatever construction we come up with. Full dependence on a company is not appropriate since a company indeed could decide to stop operations whenever it wants. However, as discussed we need better formulations.

**12:55 Joe Padfield (NGL):** *Complexity is great for building a great system, but if the basics are not understandable by non techy people then getting buy in from management will be harder, common problem though I guess.*

**Peter:** Right, to prevent the tower of bable for data science we need a solid and simple basis without mappings etc. since mappings are expensive to maintain. For me Handles can take this role, it is conceptually simple and only has two layers to maintain - a global root and a local layer. It has the independence, power and capacity to play a global role.

**13:00 Joe Padfield (NGL):** *relevance to specific sectors needs to be examined so that smaller fields can easily generate and manage their own pids, but still be bale to join in with the larger picture*

**Peter:** Right, with the Handle System one has a global root layer and so called MPAs (root nodes) form already now a global redundant network which the user does not see. Below the root node their could be services providers offering services for industry, for large institutions, for smaller institutions etc. Currently, the big nations all want their own global root node, but there is a trend that also "smaller" countries want to establish such root nodes.

**13:03Joe Padfield (NGL):** *Would one have a PID pointing at an actual data set and then another pointing at the meta data describing the data set? When does the semantic description of a data object become a data object in their own right ?*

**13:06Maggie Hellström:** *@Joe: it is certainly possible to give a PID also to the metadata - and this is implied if one follows the "PID-centric" approach .*

**Peter:** That is exactly what increassing number of communities do: have a PID pointing to the metadata since it is a DO and another one to the bit sequences (data, software, etc.) and use the PID record to link the two. A bidirectional link is established when the metadata record also has the PID of the bit sequence included.

**13:06 Joe Padfield (NGL):** *PIDs systems should be desinged to last as long as the data objects they are describing, and they always need to have suficient documentation so that migration is always possible if required*

**13:08 Joe Padfield (NGL):** *Historic data may stop being examined for its scientific data, but ideally it should be kept to allow the study of how people explored a given area of science over time. In order to understand the scientific statements made at any given time we ideally want to maintain access to the state of the knowledge and data that was avilable at the time*

**Peter:** Guess that the PID system indeed needs to be maintained for a long time period to come to a stable data domain. Repositories may come and go and one never can be sure whether a DO will survive all changes despite replications and whether the locations of the replications are known. More under B.

**13:17 Joe Padfield (NGL):** *If people need to reference small sections of a data set we would ideally have PIDs for the small sets*

**13:17 Joe Padfield (NGL):** *When we ae moving into a place were PIDs are normal, we would need to give almost everything a PID ???*

**13:21 Joe Padfield (NGL)>** *We just need to agree on the meaning of everything :-) Individual pixels in an image, individual characters on a page in a book, etc Need to have a realistic balance I supose*

**Peter:** That's what some communities advise to do - no long discussions just give all registered DOs a PID. The number of PIDs one can create is "infinite" just a few bytes. Better to create collections from atoms which allow re-combinations. However, as we have seen from Maggies comment this advice is perhaps not always appropriate. Please look also at the granularity part of the Bratislava note.

**13:20 Wolfgang Kuchinke:** *UDUS Shouldn't it be that PIDs are for everything, everybody, forever and global?*

**13:27 Joe Padfield (NGL):** *Almost every thing people work with now gets given some form of an ID, I think it would be good as part of these use cases to identify when and local ID is needed and when a PID is needed*

**Peter:** Here I can only agree. PIDs are ubiquitous. And this is the reason why we must have ONE global system we can rely on. With system I do not mean technological implementations although compliance with the specifications is crucial to prevent the tower of bable.

**13:31 Joe Padfield (NGL):** *Would this be replacing function names with PIDs?*

**13:34 Carlo Maria Zwölf:***There is also an Interest Group in RDA about data reproducibility. They are adressing similar issues on pipelines and chains of treatement. Should connect with them*

**Peter:** Indeed PIDs could be used in workflows. There is indeed a data reproducibility group, they were not so active recently ass far I understood, but perhaps it would be good to include them in the new context of the brokering and data fabric group that want to study in detail the mechanisms needed at P9 in Barcelona. I will contact Bernard and Victoria to get them interested.

**13:38 Joe Padfield (NGL):** *Does this replated to returning the ontology used to describe your data ?*

**13:40 Joe Padfield (NGL):** *OK, this is for update and or version checking, ok thanks*

**Peter:** Indeed - state information is just what some people call "system metadata" but one would not expect information such as specified by the doman ontology. The session in Barcelona will hopefully give more clarity.

**13:50 Antonio Rosato:** *The deletion option will be quite relevant within the Life Sciences.*

**Peter:** As discussed this will need some more thoughts: what to do when a DO has been deleted and only the PID record survives, what when the PID and the metadata records survive etc.