BOF on remaining PID challenges
— a Data Fabric IG “spin-off”

Peter Wittenburg, Maggie Hellström, Rebecca Grant & Carlo Zwölf
for the PID Focus Area group of GEDE
(Group of European Data Experts)

Presented at RDA P9, Barcelona April 5-7, 2017
Topics for today

• Background

• PID basics
• Granularity and collection building
• When to assign PIDs
• Versioning and PID binding role
• PID Attributes and Semantic Categories

• Usage examples from China (Lisa Liu from CAS)

• Wrap up
Background

GEDE (Group of European Data Experts)
• is a group of ca 70 European data professionals representing research infrastructures, e-infra-structures and European co-chairs of RDA Groups
• aims to promote, foster and drive the discussions and consensus forming on creating guidelines, core components and concrete data fabric configuration building
• started in July 2016, and meets regularly (telcos & F2F)
• is organized around Focus Areas groups, starting with PID usage and gathering of PID assertions
• has a web site at https://rd-alliance.org/groups/gede-group-european-data-experts-rda
Some basics

PID: persistent (and unique) digital identifier

- Handle (handle.net) is widely used PID technology
- Handle PID: `<prefix><delimiter><suffix>`
  - `<prefix>` given to registration authority and are globally unique
  - `<suffix>` is locally unique
  - `<delimiter>` is “/” for Handles/DOIs
- Examples of Handle-type PIDs are DOIs issued by DataCite and ePIC PIDs from the European PID consortium
- PIDs become actionable when resolved by e.g. handle.net: https://hdl.handle.net/11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f
- PIDs point to the location of the digital object or to a landing page
Granularity & choice of PID

• digital objects (DOs)* will be re-used and re-combined by others and we cannot predict how these objects will be used in a few years - this requires to give each scientifically meaningful object an identifier

• DOs are not just referenced within publications, but increasingly often we will need stable references for our data processing (workflows, etc.) to guarantee reproducibility

• there will be different strategies dependent on the discipline, the repositories storing data need to make their strategy clear

• there seems to be a trend that people start assigning Handles at high granularity and DOIs for citable collections (climate modelling, linguistics, etc.)

* Digital objects can be data, documents, software, media files, ...
Collection building

- in some labs it is already common practice to create **virtual collections** which are just some metadata and a whole set of PIDs pointing to DOs; **collections themselves get assigned a PID**
- Collection approach could also be used to group output from search queries (compare recommendations of Dynamic Data Citation WG)
- The Data Collection WG is working on a data model and an API standard that can be used to register and manage collections
When to assign PIDs

• for some digital content it is obvious that they are subject to changes, therefore the question is raised when (small versus major changes) one should assign a new PID to a changed object

• in some communities people work on such DOs and carry out many changes without “registering” a new version so that it can be accessed etc.

• possibly the use of versionable databases in conjunction with assigning PIDs to queries - as already suggested by an RDA working group - can address this issue, but not all communities feel this is practical or implementable

• also in this case the repositories and/or communities need to indicate which policies they follow.
When to assign PIDs (cont’d)

• in some cases it may even be useful to assign PIDs before uploading content into a repository - however then problems may occur (what about relevance and accessibility of data on notebooks etc.)

• It may help to define the term "repository" as something "simple": a "repository" is an entity whose primary tasks are to provide services to access digital object content and essential state information, given an object’s PID, and to enable reliable and trusted data management.
Versioning

• some repositories use an attribute in the PID record to refer to the **previous and/or subsequent version**; if these attributes are typed also **machines** can use the information

• other repositories use **metadata records** to include this information which is probably not as efficient as using the PID record
it is obvious that we are increasingly dependent on PIDs - thus we need to work towards a stable system that is well maintained, redundant etc.

if we have such a system we can use the PIDs to bind various types of information (bit sequences, metadata of different types, landing pages, etc.)
PID Attributes

• it is about defining a set of types, but there is no obligation to use them all
• it is generally agreed that one should not overload the PID record
• some use fragment indicators – they are not part of the PID
Semantic Categories

• there is a need for using **Persistent Identifiers for referring to concepts** and/or categories used in specific disciplines.
• it is not obvious **which kind of references** should be used to refer to semantic categories
• the **semantic web community** suggests to use **cool URIs**
• there are **existing practices** in the communities which need to be respected; in biodiversity quite a number of schemes are being used, but yet not in a systematic fashion - they are looking for an overarching schema to overcome fragmentation
Usage example

Jia Liu from the Chinese Academy of Sciences (CAS)
Wrap-up & future plans

- GEDE PID focus area group will have a F2F in June/July
- We wish to finalize our documents on
  - PID usage mapping
  - PID assertions collection
- Results will be presented to community, stakeholders, policy makers