Service for Data Retrieval via Persistent Identifiers

Vasily Bunakov[1]
[1]*Scientific Computing Department, Science and Technology Facilities Council, Harwell Oxfordshire, United Kingdom*
vasily.bunakov@stfc.ac.uk

Keywords: Data Citation, Data Retrieval, Persistent Data Identifiers, Research Infrastructures

Abstract: Persistent identifiers for research data citation have become commonplace yet current practices of minting them need evaluation to see how the data cited can be actually discovered, contextualized and processed in scalable eInfrastructures that serve both human users and machine agents. The existing means of data identifiers dereferencing can be used for basic data contextualization but more sophisticated contextualization services are required to make data readily available for automated retrieval and processing. This work takes a look at the data identifiers minting practices and discusses a possible design of a service for the machine-assisted or fully automated retrieval of formally citeable data.

1 INTRODUCTION

The recent cooperative report of four major Cluster Projects contributing to ESFRI (European Strategy Forum on Research Infrastructures) indicated that data identity is an important common topic of interest (Field et al, 2013). Minting persistent data identifiers has become a routine task in many research organizations; the use of data identifiers as references in research papers and data journals is getting popular. There are well-developed recommendations on data citation from publications (Ball and Duke, 2012) and on data promotion in global citation network, e.g. through Data Citation Index from Thomson Reuters.

However, the actual practices of persistent data identifiers assignment significantly vary across disciplines and organizations, and so does the configuration of data artefacts that identifiers designate. This makes it hard to reasonably automate the process of data retrieval which is inevitably required if we speak of scalable data infrastructures exemplified by such initiatives as EUROPEANA www.europeana.eu or EUDAT www.eudat.eu that stretch beyond the boundaries of a single data provider or a single discipline.

The level and the flavour of data openness behind persistent identifiers vary, too; machine agents of a scalable data infrastructure require meaningful structured descriptions of both non-technical aspects of access to data, such us licences and other regulation, and technical aspects of it such as APIs capabilities or the machine-executable protocols that allow data retrieval and feeding it for further processing.

For regulation aspects of machine-assisted data reuse, there have been analysis and test services provided by a few European projects and business initiatives; see in (Bunakov and Jeffery, 2013), also under Media Mixer www.mediamixer.eu and Linked Content Coalition www.linkedcontentcoalition.org. This work is going to contribute to the less explored area of modelling data APIs and data retrieval protocols, by considering one particular use case: dereferencing a persistent data identifier (that is in fact a specific API call with one parameter) with the purpose of data retrieval.

The general case for the data retrieval service using PIDs will be: a human user or a machine agent supplies a data PID to the service; the service, in case the PID resolution actually leads to the data assets, responds with the data and with contextual information (metadata) that is enough for a requester to render or analyze the data.

This work, first, considers existing effort of building data retrieval services using persistent identifiers, it then presents business analysis of the actual practices of data identifiers assignment, then this analysis is used to suggest a design of a new service for the machine-assisted retrieval of published data via persistent identifiers.

2 EXISTING METHODS OF DATA RETRIEVAL VIA PERSISTENT IDENTIFIERS

The opportunity of data PIDs use for data retrieval is well understood by some eInfrastructures yet owing to their generic nature, eInfrastructures are often more concerned about a mere incorporation of various types of PIDs and leave the care of sensible data PIDs contextualization, including for the purpose of data retrieval, to the data centres that mint PIDs (Blanke et al., 2011).

Some disciplines with an established tradition of systematic data citation in research papers, notably chemistry, have come to realize that in place of merely citing data, there are technological opportunities to get – and visualize – cited data within publications. (Harvey et al, 2015) consider three mechanisms of data retrieval via persistent identifiers:

- **10320/loc** which is a handle type that was introduced in Handle system www.handle.net to improve the selection of specific resource URLs and to add features to the handle-to-URL resolution. This mechanism further detailed in (Harvey et al, 2014) allows attaching a certain parameter to the URL upon the handle resolution; the parameter may e.g. refer to a MIME-type so the Internet browser receiving this URL knows which software application to launch in order to render the Internet resource. The limitation of this mechanism is that it relies on a specific feature of the Handle system, so only data PIDs that use Handle for resolution can be contextualized and interpreted this way.
- **DataCite Media API** that allows to associate MIME types with additional URLs as key:value pairs, so that instead of redirecting to the usual landing page, DOIs can resolve to these alternative URLs through HTTP content negotiation. The limitations of this approach are that it is vendor-specific (only PIDs minted with DataCite can use it), also it will not work if a dataset contains more than one file of the same MIME type.
- OAI-ORE Resource Maps exposed through **DataCite metadata using HasMetadata relation type**. This mechanism has been also considered by (Zenk-Möltgen, 2014) who suggested using IsMetadataFor or HasMetadata tags to refer to the richer PID descriptions (potentially suitable for the automated machine reasoning and data retrieval). These optional tags, however, a) are proprietary for a particular PID service provider – DataCite in this case, b) lack clear semantics so one will need to additionally explain it to a machine agent that say IsMetadataFor should be used for a specific sort of PID interpretation.

Alternatively, (Van de Sompel, 2014) suggested using OAI-ORE Resource Maps retrieved via “cool URIs” constructed from data PIDs according to the registered info URI scheme (Van de Sompel et al., 2006). The advantage of this approach is that it is vendor-agnostic: any PID minted by anyone can be potentially registered as an URI pointing to OAI-ORE description. The limitation of using OAI-ORE is that it is suitable indeed for the descriptions of complex data aggregations, yet may not be universally adopted by all data centres that mint data PIDs. Also, OAI-ORE lacks some features required in real practice of data retrieval via PIDs, e.g. the need, in some cases, to get authenticated or perform other actions in a certain data management system in order to actually retrieve data referred by a PID.

OAI-ORE can provide rich *descriptions* of information resources when what is actually required, if we consider the variety of data PID minting practices, is a *protocol*, or a number of protocols for data retrieval that are potentially specific to the method of how PID is associated with data artefacts.

This work is going to analyse the actual data centres practices in an open world paradigm when anyone can mint a data PID in whatever way they like. Then a generic service is suggested that utilizes this bottom-up analysis, instead of imposing a universal metadata model with substantial operational overheads for its development, adoption and maintenance across diverse data providers.

3 WHAT DATA IDENTIFIERS ACTUALLY REFER TO

There are different models and services for data persistent identifiers: Archival Resource Keys (ARK), Digital Object Identifiers (DOI), other Handle-based or otherwise designed services. For the initial analysis of data PID minting practices, the popular DataCite service www.datacite.org was looked into; it is in use by many research centres across the globe and is built upon the technical infrastructure of the Handle System www.handle.net The data providers contributing to this service should follow DataCite policy and recommendations; all these providers are deemed to be quality ones and do have the skilled staff assigned to the task of minting persistent data identifiers (which are DOIs) and data descriptions. That is why the observations on the variety of data DOI assignment practices in DataCite that we are going to communicate here reflect the diverse nature of research data rather than frivolous deviations from data curation recommendations. The information practitioners (data librarians and data archivists) just do what is appropriate for data publication in their respective research domains – which explains a good variety of information context associated with DOIs.

For the initial analysis, we randomly selected 20 DataCite DOIs minted by 16 different datacentres and looked into the following characteristics:

- **DOI default resolution targets** when DOI is resolved by a Web browser sending a standard text/html GET request. It is a responsibility of the data centre to define such a target which can be a DOI "landing" page with further links on it, or it can be something else addressable via HTTP request, e.g. a data file
- **A type of intellectual entity** referred by DOI. It is not necessarily numeric data; it may be images or other intellectual entities relevant to particular scholar discourse.
- **Data format** that defines the range of software applications for data rendering or analysis.
- **A number of "clicks" (requests)** required to get to data from DOI resolution target. "Data" here may mean various information artefacts, with various meanings and in various formats yet it is pretty clear in most cases what information artefacts the data publisher wanted to make reachable through DOI.
- **Cardinality of links to data artefacts** from DOI resolution target. DOI can be assigned to a single artefact, or a collection of them; practices of it vary across data publishers and types of data.
- **Openness of access to data**; whether the actual getting hold of data requires any form of authentication or signature (e.g. having agreed to specific Terms and Conditions of data reuse).

The initial findings are summarized in the following tables and can be a methodological foundation for further analysis when required, and for discovery of data publication patterns.

Table 1. DOI default resolution (dereferencing) targets.

DOI resolution target	Number of cases
Web page	13
MS Excel file	3
PDF file	3
XML file	1

Table 2. Intellectual entities referenced by DOIs.

Intellectual entity type of the DOI resolution target	Number of cases
Experiment, measurement or observation with numeric data as resulting artefacts not necessarily associated with any research paper	5
Numeric data associated with a research paper	3
Image	3
Research paper, report, study or PhD thesis (full text, perhaps with some numeric data in it)	7
Abstract (a short descriptions of study, no full text)	2

Table 3. Formats of data artefacts referenced by DOIs.

Data format	Number of artefacts
MS Excel	4
CSV or TAB delimited	4
PDF	6
Plain TXT	2
HTML	2
JPEG	48
CIF (crystallography data)	1
MS Word	1

Table 4. Number of HTTP requests required to get to data from DOI default resolution target.

How many requests are required	Number of cases
No click (DOI resolves directly in data)	8
One click (from the DOI landing page)	12

Table 5. Number of data artefacts published through a single DOI.

How many data artefacts are linked from DOI resolution target	Number of cases
1	17
2	1
4	1

45	1

Table 6. Openness of data access.

Whether authentication or signature is required to retrieve the data artefacts	Number of cases
No	19
Yes	1

The variations in what is offered for citing as “data” are probably most important from the information modelling perspective. Many “data” PIDs point at the descriptions of experiments or events (e.g. earthquakes), or at the full texts that represent the research discourse: doctoral theses, reports, etc. This may indicate that data per se is not always considered a citable “quantum” of research discourse – unlike research papers or detailed descriptions of experiments – so data can be sensibly cited only as artefacts of some research activity: a study, an experiment, or an observation. Some data centres minting “data” PIDs assign them exclusively to the intellectual entities that circulate in their research domain rather than directly to data artefacts, as explained by (Bunakov, 2014).

Of course, the analysis performed in this work is a small scale and may not reflect the full range of data PID minting practices across all data centres, or the actual popularity of the particular practices. As an example, the requirement of authentication or agreement with Terms and Conditions prior to getting the actual access to data (see Table 6) may be in fact more common; some data centres do require this for all or for the majority of data accessible through the PIDs.

Yet even this initial analysis suggests that the notion of “data” significantly varies across data centres and particular data cases. Also different are data formats, the cardinality of data artefacts that PIDs designate, as well as regimes and paths of access to data. In short, the *context* of data PIDs once they have been dereferenced is different so the *protocols* of data retrieval based on PIDs dereferencing should be inevitably different, too.

As an example, depending on a MIME-type of the PID resolution target (Table 1), a rendering software agent can be chosen and launched. The agent selection can be of course preconfigured in the agent’s operating system yet one may need to override the OS settings, or define specific agents for rare data formats that are not as mainstream as HTML or MS Office formats. Also, Table 4 suggests that more often than not a data artefact is not an immediate result of a PID resolution; so some protocol is required indeed in order to reach data artefacts via PIDs, like “if the PID resolution target is a data artefact, then identify its MIME-type and launch a data rendering/visualization software agent straight away; otherwise if PID is resolved into an HTML landing page, dig into it and uncover links leading to data artefacts, then get them”.

In fact, quite different protocols may be required for PIDs minted by different data centres, or for different types of data artefacts (or aggregations of them published under the same PID). The methods of how one discovers certain patterns of PIDs assignment and creates such data retrieval protocols may be different, too: as an example, one may employ text mining techniques against PID landing pages for discovering links to the actual data artefacts, or one may rely on structured annotations made by human experts about what is the path to data artefacts for a particular PID, or there may be a sustainable pattern for the context of PIDs minted by a particular data centre, so that machine calls for data retrieval can be reliably constructed on-the-fly.

What is suggested next is a principal design of a service that can support the automated construction of data retrieval protocols that are based on multiple semantic annotations submitted in a common repository by either machine agents or humans, or by a partnership of both.

4 VENDOR-NEUTRAL DATA PID CONTEXTUALIZATION SERVICE

The limitations of the existing solutions mentioned in Section 2 of this work and the notion of data retrieval protocol introduced in Section 3 suggest that a universal service for data PID contextualization that runs independently from any of the existing PID service providers should be viable. This new service may not require indeed all the PID resolution mechanisms that existing data PID service providers are offering; what it will need from them is only a PID itself which can be associated then with as rich contextual descriptions as required.

Allowing and registering sensible statements about data PIDs, made by data curators from data centres or by third parties – e.g. by researchers who produced the data, or by machine agents (employing text mining, machine learning or other techniques) could be indeed a mechanism for collaborative curation of data PIDs context. The examples of granular statements about data PID context expressed in plain English will be:

‘X is a data PID minted by data centre Y’

’X designates the full text of a doctoral thesis according to human agent Z who made a statement about it’

’X resolves in a PDF file having URL www.xxx.url according to machine agent M that made a check on DD-MM-YYYY at HH:MM:SS’

’I trust data centre Y and human agent Z, also believe the machine agent M is well tested and untapped’

From which statements, once they have been registered and shared, someone (that may be a machine agent / reasoner) can derive that in order to render data behind the PID X, the PDF compatible visualization software is required and that this data PID in fact represents the full text of a doctoral thesis. Also the executable description of the data artefact can be produced in order to actually retrieve it (via its URL in this case). Notably, the reasoning is performed in a situation when statements about the data PID could have been independently made by various agents in the “open world” paradigm when anyone can be, to a certain extent, a PID context curator. Filtering / selection of particular statements for reasoning over them is a responsibility of a sensible PID context resolver that could be a human, a machine agent, or a partnership of the two.

A generic protocol for machine-assisted data contextualization and data retrieval via persistent identifiers dereferencing may look then as follows:

1. The agent resolves the data PID and tries to uncover a number of aspects: what type of intellectual entity the PID designates; how many data artefacts are there, and what are their formats; whether any authentication or signature is required for data retrieval; what is the path (or the sequence of requests required) to the data artefacts associated with the PID.
2. The agent makes some sort of automated inference on the above aspects or/and requests the opinion of a human user which options to select.
3. The agent forms the request to get the data artefacts and their context/metadata (as advised by the previous step), and arranges for the authentication or the signature if required.
4. The agent looks for a software application suitable for data rendering or analysis, and feeds the data artefacts and their context into that application.

The existing data PID management services such as DataCite will be involved in this new service only initially when they assist in minting PIDs. Everything else: collecting statements about PIDs, reasoning over statements, and actions resulted from reasoning can be supported by a new vendor-agnostic service independent of existing PID management providers.

The schematic view of this new independent data PID contextualization service aimed at the automated data retrieval using data PIDs is presented in Figure 1:

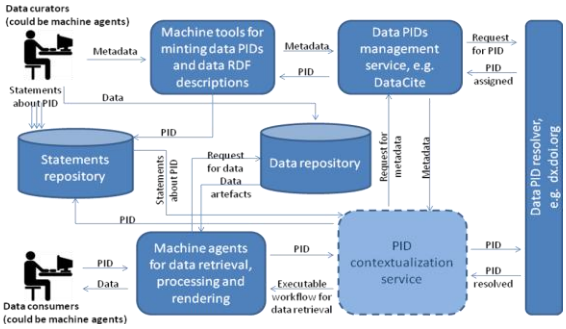
Data curators
(could be machine agents)
Metadata
Machine tools for minting data PIDs and data RDF descriptions
Metadata
PID
Data PIDs management service, e.g. DataCite
Request for PID
PID assigned
Statements about PID
Data
PID
Statements repository
Data repository
Statements about PID
Request for data
Data artefacts
Request for metadata
Metadata
Data PID resolver, e.g. dx.doi.org
PID
Machine agents for data retrieval, processing and rendering
PID
PID contextualization service
PID
PID resolved
PID
Data
Executable workflow for data retrieval
Data consumers
(could be machine agents)

Figure 1: Suggested service design.

One of the advantages of this service is that it will allow the automated data retrieval using all sorts of PIDs: DOIs, ARKs, and bespoke identifiers, as the statements can be made and shared in a common repository for any of the existing PID types.

From technology perspective, the statements repository could be based on a massive graph database or a triple store, or a federation of them, and on the commonly accepted, vendor-neutral (yet, potentially, domain-specific) definitions of the executable data retrieval protocols that result from fully automated or machine-assisted reasoning over shared statements about data PIDs.

A pilot service could be built first for a particular research community, then scaled up in a multi-domain environment, which will prove then the universality of the approach suggested. Facilities science with its established data acquisition and data sharing practices outlined in (Bunakov et al., 2015) can be a perfect case for such a pilot. There is an ongoing effort within PanData initiative www.pan-data.eu of building a machine-interpretable description (ontology) of facilities science domain, and EUDAT project www.eudat.eu performs experiments on the scalability of RDF triple stores and graph databases that should be able to manipulate substantial numbers of granular statements about data PIDs. These two streams of work can be joined for building a working prototype of a scalable service in support of data PIDs contextualization and automated data retrieval.

5 CONCLUSIONS

A variety of practices for minting data persistent identifiers leads to disparate and semantically inept representations of data PIDs context that makes it difficult to use the existing PID resolution mechanisms for automated data retrieval.

This work, first, considered existing suggestions for IT architecture in support of data retrieval via PIDs. Secondly, it suggested a methodology for the analysis of data PID minting practices to be taken into account for the design of an automated data retrieval service, and presented an example of such analysis. Thirdly, a particular design of data retrieval service was suggested, based on data PIDs semantic annotations (statements) shared in a common repository, perhaps underpinned by a federated infrastructure. Also, a potential for a vendor-neutral implementation of a pilot service in a certain data publishing domain was indicated.

This work is a contribution to business analysis and IT architecture required for such a service and should help to support the implementation of it.

ACKNOWLEDGEMENTS

This work is supported by EDUDAT project www.eudat.eu funded by the EU Horizon2020 research and innovation programme. The author thanks his colleagues in EUDAT for the discussions, although all the views presented are those of the author and not necessarily of the EUDAT collaboration.

REFERENCES

Ball,A., Duke, M., 2012. *How to Cite Datasets and Link to Publications.* Digital Curation Centre, 2012

Blanke, T., Bryant, M., Hedges, M., Aschenbrenner, A., Priddy, M., 2011. Preparing DARIAH. In *IEEE 7th International Conference on E-Science (e-Science), 2011.*

Bunakov,V., 2014. Investigation as a member of research discourse. In *RCDL 2014 – 16th All-Russian Conference on Digital Libraries. Dubna, 2014.*

Bunakov, V., Jones, C., Matthews, B., 2015. Towards the Interoperable Data Environment for Facilities Science. doi:10.4018/978-1-4666-6567-5.ch007. In *Collaborative Knowledge in Scientific Research Networks, chapter 7, 127-153. IGI Global, 2015.*

Bunakov, V., Jeffery, K., 2013. Licence management for Public Sector Information. In *2013 Conference for E-Democracy and Open Government (CEDEM'13), Krems, Austria, Edition Donau-Universität Krems, 2013*

Field, L., Suhr, S., Ison, J., Los, W., Wittenburg, P., Broeder, D., Hardisty, A., Repo, S., Jenkinson, A., 2013. Realizing the full potential of research data: common challenges in data management, sharing and integration across scientific disciplines. In *e-infrastructures User Forum, CERN, 18-19 November 2013.*

Harvey, M., Mason, N., McLean, A., Rzepa, H., 2015. DOI-2-data: Interoperability for Data Repositories. Metadata-based procedures for Retrieving Data for Display or Mining Utilising Persistent (data-DOI) Identifiers. *Retrieved from http://www.ch.ic.ac.uk/rzepa/mason/rdm/DOI-to-data.html*

Harvey, M., Mason, N., Rzepa, H., 2014. Digital data repositories in chemistry and their integration with journals and electronic notebooks, *J. Chem. Inf. Mod., 2014, 54, 2627-2635. doi:10.1021/ci500302p*

Van de Sompel, H., Hammond, T., Neylon, E., & Weibel, S., 2006. The "info" URI scheme for information assets with identifiers in public namespaces (Network Working Group Memo RFC 4452). *Retrieved from http://tools.ietf.org/html/rfc4452*

Van de Sompel, H., Sanderson, R., Shankar, H., Klein, M., 2014. Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. *International Journal of Digital Curation. Vol. 9, Iss. 1, 331–342, doi: 10.2218/ijdc.v9i1.320*

Zenk-Möltgen, W. Machine Actionable Integration of DataCite and DDI Metadata. In *EDDI14 – 6th Annual European DDI User Conference. London, 2014.*