German Data Forum
(RatSWD)

www.germandataforum.de

RatSWD
Working Paper Series

Working Paper No. 147

What Makes Persistent Identifiers Persistent?

Nikos Askitas

June 2010

Federal Ministry
of Education
and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: German Data Forum (RatSWD) | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

What Makes Persistent Identifiers Persistent?*

Nikos Askitas†

IDSC of IZA

June 28, 2010

Abstract

This essay sketches technical and non-technical issues around persistent identifiers (henceforth PIs) in a manner which makes no attempt to be complete. Our goal is to rescue the core notions from the obscurity which detail and completeness burdens them with. A reader willing to entertain the idea of their necessity should be able to cut to the chase and follow a more complex and involved debate after reading this. Our hope is that in isolating the core issues we will enable a more founded discussion of the social and political issues involved in PIs.

Keywords: Persistent Identifier, handle, DOI, data, trust, URN, RePEc

*This note is a slightly modified version of a presentation I have given to the 2nd meeting of the RDC and SDC's of the German Council for Social and Economic Data which took place at the DIW in Berlin on June 14 2010.

†Address: Schaumburg-Lippe-Str. 5, D-53113, Bonn, Germany, telephone: +49-228-3894-525, e-mail: *askitas@iza.org*

1 What are persistent Identifiers all about?

With the advent and the proliferation of electronic publishing on the web a large population of "objects" (e.g. documents, qualitative or quantitative data etc) has been amassed, one of the main characteristics of which is its volatility: "objects" (i.e. documents, data etc) may no longer live at the URL they were once seen making their citation a futile business. This is obviously an undesirable state of affairs. Persistent Identifiers are an attempt to solve this problem. It is important to be mentioned here that the volatility of the aforementioned population of objects consists not only of the fact that "links die" but also that they are "born" in great numbers thus imposing a requirement of scale and volume on potential solutions. Persistency in this context is a concept which is not served by the fact that a multitude of solutions exist as this would defeat the purpose: having a race between several approaches is good for choosing the right thing but inspires no high hopes of persistency since some of the candidates are bound to disappear taking the documents' persistency to the grave. This implies that choice considerations should be made early on before investing in any one direction and once they have been made a solution should be chosen which either covers the specified needs or can evolve to cover the rest. The parallel existence of solutions under the requirement that they are compatible with each other in some way is not unthinkable.

To complicate matters a bit what we want to consider here is persistent identifiers for data (mainly quantitative data but this is to a large extend irrelevant for our essay as the "type" enters the picture at the metadata level which we do not handle here). In order to motivate the need for PIs in the area of quantitative datasets we need to discuss the issue of citing datasets in empirical research. A good attempt to open this discussion as well as recommended reading is [AK2007], where the question of how to cite datasets is being discussed and persistent identifiers come into the picture. To summarize this problem suppose that you have a body of empirical research literature (say economics papers) and you want to compile statistics regarding the data its various papers use. Being able to do so in an easy, machine actionable way would be beneficial to all stakeholders of the research lifecycle including the data producer. You could for example have a mapping between JEL codes and datasets formalizing the impact (and its scientific distribution) of data on research areas and topics. Since there is no practiced formal citation of papers, analyzing this problem is a cumbersome task. The difficulty in doing this analysis prevents the data producers from getting a good handle on the use of their data. The researchers would also benefit from such an analysis as they would be able to easily answer the question: which datasets have been used (and how extensively) in the JEL codes of their interest.

At the IDSC of IZA there is ongoing work which attempts to solve this problem for the IZA Discussion Paper Series. It is obvious from this discussion that data citation cannot begin to happen before you have some persistent way to refer to them: it is of no use citing an object whose existence is not persistent. This sufficiently motivates the rest of this essay and sets the stage for what follows. We will try to formalize the discussion by reducing it to its essence and get to the core of persistence technologies. We will discuss the importance but also insufficiency of technology to solve the problem and bring the rest of the necessary ingredients into the picture.

2 The central idea and its consequences

We want to clarify the central idea of the technical aspect of PIs. In Figure 1 we depict symbolically a population of observers (which we may think as citations) and a population of objects (which we may think as publications). A change in the location of any of the objects causes us to have to maintain **all** citations of that object.

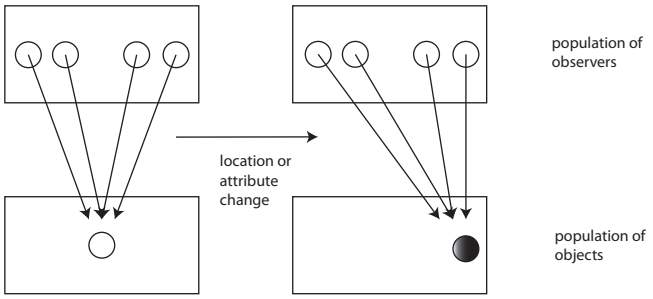

Figure 1: A location and/or attribute change in the population of objects causes the need to notify the entire population of observers for every object changed

A technical and easy way to avoid this is to introduce an abstraction layer between the objects and their observers as in Figure 2.

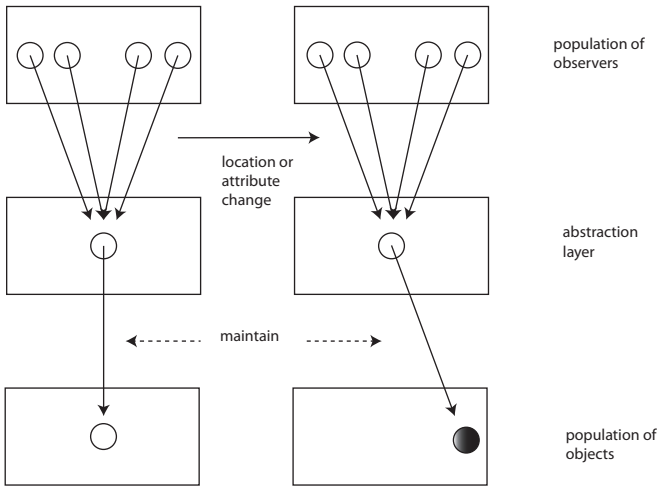

Figure 2: Inserting an abstraction layer between the objects and their observers.

In doing so we achieve two things. First we reduce the amount of maintenance by a factor equal to the number of observers involved and second observation becomes "canonical" now having a chance to thus be robust i.e. persistent. In Figure 1 changing an object or any of its attributes caused us to need to change all arrows pointing to it from all observers. Now we just need to change one arrow. This is progress indeed

but only because we make a silent assumption: We assume that the abstraction layer does not change. For the reader who likes to think in analogies the simplest kind of such abstraction layer is a Shortcut on ones desktop: Changing the version of the underlying program does not break the functionality of the shortcut. You may run another program but the functionality will be intact.

Now let us spell out the four major consequences of such an abstraction layer:

C1. The continuity i.e. persistency of the observation is NOT achieved if one does not maintain the mapping between the abstraction layer and the population of objects otherwise you still end up with a lot of "dead links".

C2. The continuity i.e. persistency of the observation evaporates all at once if the abstraction layer ceases to function or be present.

C3. The abstraction layer can but should not hold many pointers to one object.

C4. The abstraction layer is aware of ALL observation.

In what follows we will discuss these consequences and translate them in the context of citing data from publications. Before we do so we would like to point out that none of these consequences are necessarily negative or positive in and within themselves but one needs to understand them and make informed choices.

The first consequence (labeled with C1) is substantial. While there is now "less" to maintain the impact becomes higher: Failing to maintain one arrow between abstraction layer and object layer impacts multiple observers all at once.

C2 is basically an extreme form of that as it invalidates all observations at once. This would be a citational armageddon so to speak. It is a misunderstanding that the distributed nature of systems can prevent that from happening. The "root" servers are sine qua non[1].

Consequence C3 is important. Many pointers to the same object leads to the usage of many names for that object and defeats the purpose of exhaustively knowing object citations. This is best seen in Figure 3 in the case of many providers.

Finally C4 looks innocent but it isn't. The abstraction layer knows every observation from any observer to any object. It alone has this privilege. Regulation is needed here to determine who the data owner is, what the disclosure degree of the data is etc. Figure 4 helps put a "price tag" on the consequences C1, C2, C3, C4 a bit more context specific. If for some reason the resolver of the middle (abstraction) layer in Figure 2 is absent then the right hand side of Figure 4 is a large citational cemetery. If any of these PI are not maintained then they are as dead as the corresponding link on the left of Figure 4.

One of the technical candidates for implementing PIs for data is the `handle.net` flavor (to be more precise this is just a system of canonical naming which may be used as an ingredient in building Object Persistency). A special variant of that is the `doi.org` variety which does just that: uses the handle system as a technical foundation. In what follows we want to recast the discussion above using the DOI context as the backdrop against which we highlight the significance of the points made.

[1]The reader can verify this in a context different but similar in nature. On May 13 2010 the root DNS servers owned by DENIC responsible for mapping domain names under the .de level to IP addresses failed to function properly bringing the "German Internet" down. The main implementations of PIs copy DNS with good reason: it is well thought out. It can fail nonetheless.

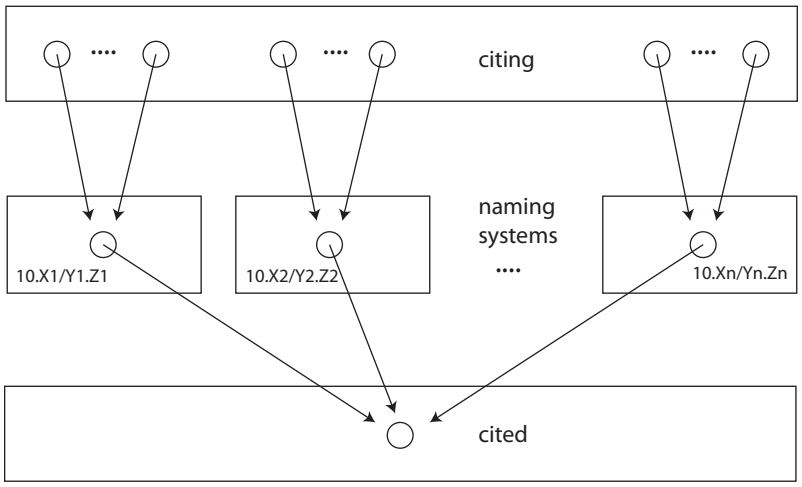

Figure 3: Having many "names" i.e. PIs for the same object is probably not a good idea. If citations use them independently we are back to square one: we still cannot know all names. This is a bad thing regardless of whether the "providers" of name i.e. PIs use the same or different technologies/solutions

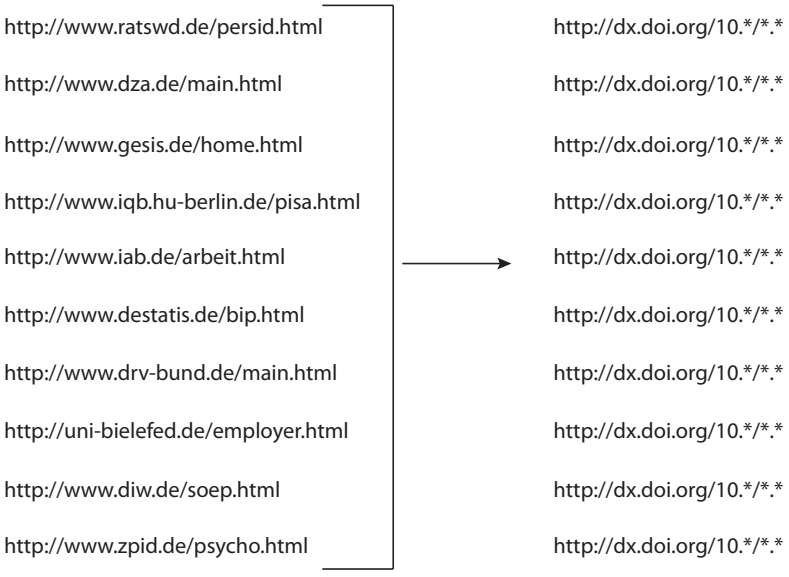

Figure 4: Sample links before and after

C3 says we should have one persistent link per object, since every additional one undoes the main benefit of PIs. This needs to be elaborated upon because it seems to be a candidate for violation in the case of DOI and its business model. A DOI has the form `doi:10.A/B.C` where the left hand side with respect to the forward slash is reserved for the so called Registration Agency and the right hand side for the object. If you are the object's owner you think of `10.A` as a Service Provider. So you should be able to take your objects elsewhere if you for some reason want to **without** breaking functionality of existing references. This is not possible though because the Service Provider's canonical name becomes part of your object's canonical name. If you want to have someone else maintain your objects you have to leave the old ones behind for ever. Otherwise you have to rename your objects but this violates persistency in more than one ways. If you now dropped your contract with `10.A` they may no longer feel obliged to maintain the old names so references to your objects may die a sudden death. This means that a different type of canonical name maybe necessary, one which is provider neutral and while at that quite likely capable of expressing domain of knowledge, context, country etc very much similar to the DNS system. This will then be a viable business model which will be sustainable and growth oriented.

C4 is most likely the one which no one wants to think about early, most people will let it fall under the table and it could come back to bite us the hardest at some point in the future. The resolvers of the service provider (in this case doi and/or handle) will know everything: which dataset gets how many hits, from which IP or geo coordinates and how does all this vary over time. This is not just about academic citations. It is about clicks. If the hits come from "observers" i.e. documents with PIs themselves you will know which research areas use which data. All this data is fascinating and quite possibly powerful. The question is: Whose data is it? Who has access to it? The DOI system is what `crossref.org` uses which is controlled by a consortium of over 3000 publishers. In the DOI nomenclature `crossref.org` is a Registration Agency. Now the fact that there is large commercial support for this type of PI attempt may be viewed as a blessing or a curse depending on your persuasions. While there is freedom of choice regarding persuasions there is no way around being confronted with these issues for those deciding adoption.

There is enough evidence out there that the commercial enterprises in the publishing industry are not particularly forthcoming with information they can put their hands on (for good reason since they deal in information) so one needs to cover all basis contractually from the outset. The middle layer is best captured by Alice's and Humpty Dumpty's dialogue which we paraphrase from L. Carroll's Alice in the Wonderland:

> *The question is,* said Alice,
> *whether you can make PIs mean so many different things.*
> *The question is,* said Humpty Dumpty,
> *which is to be master thats all.*

In the last segment of this section we construct an example using real entities we all know about for purely pedagogical reasons: in order to provide the reader with a sufficient amount of language and examples to debate intelligently and decide informed. Our objects will be Discussion Papers from the IZA Discussion Paper series. In the RePEc nomenclature this series is identified by the handle `RePEc:iza:izadps`. From this series we will use two papers in order to make a couple of points: Discussion Paper No 14 ([HWZ1998]) and Discussion Paper No 2280 ([KT2006]). We will use

two different systems of canonical naming both of which: a. have the potential to provide the basis for object persistency and b. do so to some extend or another in their own way. The systems we will use are RePEC (`www.repec.org`) and the German National Library (`https://portal.d-nb.de/`). What the systems share in common is that they allow us to create canonical names for objects in their realm and allow us to resolve the canonical names to a minimum set of metadata. In the case of the German National Library the system is a so called URN:NBN[2] based system whose resolver is `http://nbn-resolving.de/`. In the case of RePEc it is an implicit resolver i.e. all the ingredients to build it are there but it does not exist yet[3]. What is important here is to demonstrate in concrete terms some of the things discussed earlier. Now [HWZ1998] has a canonical name in RePEc and the name is `RePEc:iza:izadps:dp14`. It does not have a canonical name (yet) in the `http://nbn-resolving.de/` system. Our second object [KT2006] has a canonical name in both systems. In RePEc it is called `RePEc:iza:izadps:dp2280` and in the German National Library system it is called `urn:nbn:de:101:1-200912071035`. Now let us focus on the object [KT2006] and its two names `RePEc:iza:izadps:dp2280` and `urn:nbn:de:101:1-200912071035`. Both systems have internal reasons why they need to assign canonical names to their objects. Both systems have the right to do so since canonical naming belongs to the basic rules of good digital housekeeping. Both systems will use their own canonical naming to refer to these objects. The objects themselves though which come from sources (Publication Agencies) foreign to the two systems refer to other objects in ways they are free to choose. The moment anyone of the two systems proposes itself as not only a canonical naming system but as a Persistent Identifier system the question of referring to [KT2006] with this name: `RePEc:iza:izadps:dp2280` or that name: `urn:nbn:de:101:1-200912071035` becomes equivalent to the following question: Who do we trust?

3 Conclusions

PIs are an interesting solution to a problem we need to identify. As every solution they represent less the extinction of all problems and more a choice of the problems we want to be dealing with in the years to come. Over 200 years ago Adam Smith wrote "The wealth of nations" and over 2000 years ago Aristotle was pondering on "the art of wealth acquisition". This means that scholarly economics to mention one example survived at least 2000 years without PIs. This is not to say we do not need PIs but it serves to put things in perspective.

Persistency on a technical/digital level is nothing but the attempt to install a centralized global webmaster. There is hierarchy and distribution as well as delegation of responsibility which make the system both scalable as well as highly available (both crucial features for performance and operative reasons) but the mechanism of communication between the various entities creates in effect centralized web masters or

[2]The reader interested in learning about these abbreviations may want to take a look at `www.persid.org`. Good reading for Implementation of and Policy regarding PIs are [HK2006] and [NWB2009] respectively.

[3]Under `ftp://all.repec.org/RePEc/all/` one can find all series and archives with their location (RePEc is a bottom up system) and at every location according to some simple rules there are all the objects and their metadata. The site `http://econpapers.repec.org/` may of course be thought of as a resolver as Christian Zimmermann pointed out to me.

substitutes thereof. For users who have been around long enough another metaphor which demonstrates the essence of persistent identification technology is the by now obsolete Library Card Catalogues (Figure 5). It is as easy to construct the technology necessary to implement any flavor of PIs today as it was to construct a Card Catalogue in its time.

Figure 5: A Card catalogue of the University Library of Graz. Photo by Dr. Marcus Gossler: a system which is maintained and offered so one does not have to run around browsing shelves but a well organized system of drawers were Cards contain metadata for the objects AND its current location

The main point is that persistency is not about technology but about commitment of communities organized by knowledge domains. To quote Sun Microsystems cofounder Scott McNealy "technology has the shelf-life of a banana" so whichever way we go we need to safeguard against it all becoming obsolete tomorrow. The persistency of PIs is a pledge and a commitment one makes. It therefore benefits if it is built on community values such as trust. The plurality of communities increases chances of survival (as we know from evolution theory) whereas the fact that these communities are centered around domains of knowledge implies getting the modalities straight. Networking with other domains is vital for the same reasons. Choosing a healthy business model is vital and it is not helped by making the Service Provider's name part of the the Object's name. The upshot of the argumentation in this essay is that persistence is not about technology. Quite the contrary is true. Persistence has got to be technology neutral and be based on a pragmatic calculus whose numerator includes the community, domain knowledge and trust. Persistency is a pledge and it all therefore comes down to who you believe. You've been warned[4].

[4]At the time of this writing GESIS and the IDSC of IZA under the umbrella and sponsorship of the Council for Social and Economic Data (ratswd.de) are planning a two day workshop on the subject In February 1-2 2011 in Bonn.

References

[AK2007] M. Altman, G. King, *A Proposed Standard for the Scholarly Citation of Quantitative Data*, D-Lib Magazine, 2007, Vol. 13, Num. 3/4

[HWZ1998] R. Hauser, G.G. Wagner, K.F. Zimmermann, *Memorandum: Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestuetzter wirtschafts- und sozialpolitischer Beratung* , IZA Discussion Paper Series, 1998, No 14, RePEc:iza:izadps:dp14

[KT2006] K. Tatsiramos, *Unemployment Insurance in Europe: Unemployment Duration and Subsequent Employment Stability* , IZA Discussion Paper Series, 2006, No 2280, RePEc:iza:izadps:dp2280

[HK2006] H.-W. Hilse, J. Kothe, *Implementing Persistent Identifiers*, `http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8`, `ISBN 90-6984-508-3`

[NWB2009] N. Nicholas, N. Ward, K. Blinco, *A Policy Checklist for Enabling Persistence of Identifiers*, D-Lib Magazine, 2009, Vol. 15, Num. 1/2