# **Title:**

# **PIDs, please play FAIR and identify yourselves!**

# **Abstract**

This is an extended, revised version of . Findability and interoperability of some PIDs and their compliance with the FAIR data principles are explored, where ARKs were added in this version. It is suggested that the wide distribution and *findability* (e.g. by simple 'googling') on the internet may be as important for the usefulness of PIDs as the *resolvability* of PID URIs. This version also includes new reasoning about the low adoption rates of PIDs such as DOIs for citation. The prevalence of phenomena such as *link rot* implies that the persistence of ordinary (non PID) URIs cannot be trusted. By contrast, the well distributed, but seldom directly resolvable ISBN identifier has proved remarkably resilient, with far-reaching persistence, inherent structural meaning and good *validatability*, by means of fixed string-length, pattern-recognition, restricted character set and check digit. Examples of regular expressions used for validation of PIDs are supplied or referenced. The suggestion to add *context* and meaning to PIDs, making them "identify themselves", through namespace prefixes and object types is more elaborate in this version. Meaning can also be conferred by means of structural elements, such as well defined, restricted string patterns, that at the same time make PIDs more "validatable". Concluding this version is a generic, refined model for a PID with these properties, in which namespaces are instrumental as custodians, meaning-givers and validation schema providers. A draft example of a Schematron schema for validation of "new" PIDs in accordance with the proposed model is provided.

# **Introduction: Identifiers in science**

Identifiers in science may refer to digital or physical objects, or concepts. They may also refer to persons, researchers, such as ORCiD IDs , or to research organizations, such as the recently launched RORs . This paper will focus on persistent identifiers (PIDs) for research *outputs*, 'things' such as articles, datasets, samples, concepts etc. But, as suggested in section 7, ORCiDs or RORs may be an optional part of a modular, integrated identifier for research outputs. PIDs may be general or domain-specific. Among the more prevalent general PID-types are ARK, DOI, Handle and UUID. There are also old, bibliographic identifiers like ISBN. Created in the 1960's and 1970's of the print era, how come they survived into this digital age? Some reasons might be: they are well distributed across the internet, and widely used by stakeholders (libraries, publishers, readers). They have a semantic structure, identifying well-defined objects, and a fairly precise validation mechanism through fixed string-lengths, limited character-set and check digits. Some of these properties of good identifiers are shared by ARKs, DOIs, Handles and UUIDs, or other more domain specific identifiers used for scholarly data, but seldom all of them simultaneously. The focus here is on findability and 'validatability' of PIDs of different types.

# **Identifiers - why do we need them?**

The general purpose of identifiers is to serve as *references* to the objects that they are supposed to identify. Some identifier strings indicate what *types of objects* they are meant to identify. Far from all identifiers fulfil this requirement. It is often left to the *names* of things to provide context and *meaning*. Context may be added by means of location within an hierarchical system, e.g. as in Linnéan taxonomy, where scientific names situate a species within a genus, sometimes also containing the provenance of that name, serving to disambiguate between names of species belonging to widely different genera, e.g. *Asterina gibbosa* Gaillard 1897 (a fungus), and *Asterina gibbosa*(Pennant, 1777) (an echinoderm, a starfish). For the same purpose, it also happens that 'things', objects are renamed later, as with the preceding fungus species now having the accepted scientific name *Asterolibertia gibbosa (Gaillard) Hansf. 1949* and / or are assigned an identifier: *urn:lsid:catalogueoflife.org:taxon:02af8238-ac8f-11e3-805d-020044200006:col20150401*.. However, even if a PID may well serve the need for disambiguation by uniquely identifying an object, it may still be no better - sometimes, as in this case, perhaps even worse - at giving access to said object, or at least a landing page with metadata about it. The identifier assigned above is neither directly resolvable nor 'googlable', while the scientific name is at least easily findable via a search engine.

While scientific *names* are often useful for *describing* objects, they have other drawbacks compared to PIDs, some of which were identified by [[36]](file:///C%3A%5C%5CUsers%5C%5Cjoph9849%5C%5CDesktop%5C%5Cphilipson-savesd17revext%5C%5Cphilipson-savesd2017revext.html%22%20%5Cl%20%22patterson-2016%22%20%5Co%20%22Patterson%2C%20D.%20et%20al.%20%282016%29.%20Challenges%20with%20using%20names%20to%20link%20digital%20biodiversity%20information.%20Biodiversity%20Data%20Journal%204%3A%20e8080%20%2825%20May%202016%29.%20https%3A//doi.org/10.3897/BDJ.4.e8080). For example, homonymy and disambiguation should be no problem for globally unique identifiers. And while concatenations or abbreviations may be problematic in the use of names for identification, string-length and pattern restrictions are useful for validation of identifiers. Missing or added characters, and some types of misspellings are easier to detect and validate in standardized identifiers of fixed string-length or well-defined character patterns. However, these properties of some identifiers may conflict with the interest in having also transparent, meaningful PIDs that at least in part "speak for themselves". The result of a compromise between these two interests may be seen in the *Handle* system (below).

# **FAIR principles**

The FAIR guiding principles aim to make (meta)data **Findable, Accessible, Interoperable, and Re-usable**.As such they concern also PIDs, as is seen from some of the principles:

The FAIR data principles.

The FAIR principles clearly need interpretation to become fully operational, and such work is also well in progress. Further explications of some of the principles are also available in . Figuring prominently in the explications of all these principles, particularly *interoperability*, is the requirement that metadata should be machine readable, a *conditio sine qua non* for FAIRness.] This is also used by *fairmetrics.org* as a measure of Findability.

However, the FAIR principles do not say anything *explicitly* about *validation*. Particularly for the principles of *Interoperability* and *Re-usability*, it is crucial that metadata can be properly validated against a schema, as adhering to an accepted metadata standard. It has been remarked that this is already implied by the FAIR principle R1.3 above, but even so, only indirectly and in a way that is still open to interpretation. There are several cases where general data repositories, professing to be FAIR and to adhere to accepted metadata standards both for their default output and export formats, nevertheless fail to validate against schemas of these same standards.*Fairmetrics.org* explicates R1.3, as measuring a "Certification from a recognized body, of the resource meeting community standards", by means of a valid electronic signature, such as a verisign signature.However, one might ask whether general data repositories such as [Harvard's Dataverse](https://dataverse.harvard.edu/), [Figshare](https://figshare.com/) or [Zenodo](https://zenodo.org/), qualify as "recognized bodies" in this respect, all being part of the test reported in "Evaluation\_Of\_Metrics/Supplementary Information\_ FM Evaluation Results.pdf" , but none of which could be evaluated on R1.3. This comes as no surprise, since there is already a comment on R1.3 saying that "Such certification services may not exist, but this principle serves to encourage the community to create both the standard(s) and the verification services for those standards."True, in the rationale for FM\_R1.3 there is mention of validation: "... As such, data should be (individually) certified as being compliant, likely through some automated process (e.g. submitting the data to the community's online validation service)".But it remains unclear if the "community" here refers to a general metadata standard, or a repository using its own standard and validation service? Some output metadata files from repositories even lack a *schemaLocation* reference, making it difficult to validate them, or, the schemaLocation given might be erroneous, as observed in one case.To see if repositories are in compliance with standards, one must use available validation tools, testing to what extent they are keeping their promises of compliance. This concerns metadata in general, but naturally includes also identifiers. We must be sure that they are of the type or format they claim to be, even if they cannot be resolved in a browser. Failed validation, e.g., due to simple typos or wrong namespace, may help explain why an identifier or URI does not resolve as expected. Validation is also important for the possibility to export metadata to another format, thereby promoting the re-use of data, without exporting also potential errors. Resistance to transcription errors, e.g. by means of a restricted character set, using base32 for encoding, and fixed string-length (suffix has 2 times 4 characters, separated by a hyphen), has been promoted as an advantage of so-called "cool DOIs". These are precisely the kind of properties that make PIDs eminently "validatable" and therefore more likely to be correct. Although transformation or harvesting of metadata might be possible even without validation, trust in the results and quality as well as the eventual findability of the data (and so again the re-usability) might be seriously affected. For enhanced findability, it is also important that standard, widely distributed identifiers are used.

Validation of an identifier means ensuring that it is true to its proclaimed type, for example, making sure that what is flagged as an ISBN is not in fact an ISSN (real use case), or that the string-length and checksum is compliant with its type. An advantage of promptly validatable identifiers is that validation can often be performed off-line, by means of a pattern expressed as a regular expression, a script (JavaScript, Python, etc.), an HTML form, or an XML-editor.

# **Resolvable or findable?**

In the current FAIR principles, the focus is very much on *resolvability* of identifiers, du e to the general awareness of phenomena like 'link rot' and 'reference rot'.It has even been suggested to put up digital tombstones over disappeared resources, with metadata from their last known whereabouts serving as epitaphs.A 2013 study in BMC Bioinformatics analyzed nearly 15,000 links in abstracts from Thomson Reuters’ Web of Science citation index and found that the median lifespan of web pages was 9.3 years, and just 62% were archived.This happens although there is an understanding that [u]nique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.A recent study of some 40 research data repositories found that only one of these (3%) was compliant with the FAIR principle of Accessibility requiring a clear policy statement (or various examples of data this has actually happened to) indicating that metadata is still available even if the data is removed. The argument here is not that resolvable, persistent URIs should be avoided as identifiers, but they may not be sufficient. But, as has been eloquently remarked, "persistent URIs must be used to be persistent". Resolvable URIs as PIDs work by decoupling the location and the identification functions of URIs.

The custodian of a web resource maintains the correspondence between the identifying URI and the locating URI in the resolver’s look-up table as the resource’s location changes over time. ... The solution comes at a price because it requires operating a resolver infrastructure and maintaining the look-up table that powers it.

This is true of ARKs, DOIs, as well as Handles, PURLs and URNs. There are in fact numerous cases when the lookup-table is not maintained and updated as required. That is why it may be wise not to rely on a single 'custodian' for the resolution of identifiers and access to associated metadata. Note that we are not talking here about simply having more than one proxy server acting as resolvers of the same PIDs. We already have that; provided the lookup-table is managed properly, the three different DOI-URIs from three different proxy-servers all resolve to the same landing-page location: <https://doi.org/10.1007/978-3-319-53637-8_11>, <https://hdl.handle.net/10.1007/978-3-319-53637-8_11> and [https://identifiers.org/doi:10.1007/978-3-319-53637-8\_11](https://identifiers.org/doi%3A10.1007/978-3-319-53637-8_11). ARKs (Archival Resource Keys) are resolved by *identifiers.org* and *n2t.net*, as well as by their "mother institutions", e.g. [n2t.net/ark:/67531/metapth346793/](https://n2t.net/ark%3A/67531/metapth346793/), [identifiers.org/ark:/67531/metapth346793/](https://identifiers.org/ark%3A/67531/metapth346793/) and [digital.library.unt.edu/ark:/67531/metapth346793/](https://digital.library.unt.edu/ark%3A/67531/metapth346793/) resolve the same content. It is rather the *distribution* and *use* of identifiers - whether resolvable or not - that is important here. It seems not even the authors of [[43]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cphilipson-savesd2017revext.html#vandesompel-2016a) are true to their own principles, since three of their references that actually have DOIs are cited without them: [[9]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cphilipson-savesd2017revext.html#duerr-2011) [[29]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cphilipson-savesd2017revext.html#klein_etal-2014) [[51]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cphilipson-savesd2017revext.html#zhou-2015). So, despite having DOIs or other PIDs assigned, documents are often not cited by those PIDs. One possible reason might be that the PID is not clearly displayed in the landing page with metadata, or in the document itself. In the case of above it takes an extra click on a link 'Cite as' to actually have the DOI displayed. But that should hardly be the reason why it was not used for citation in , since the citation there is actually much more verbose and complex, than it would have been to just copy-paste from the 'Cite as' page above. Another, slightly ironic case concerns , the founding paper of the FAIR principles, where you either have to download the citation with the DOI from the landing page, where it is not displayed, or find it at the bottom of each page in the actual paper, but not prominently marked. This may partly explain why a recent paper on software sustainability and reproducibility, while arguing that one of the ways to make software more reproducible is to "use a persistent identifier such as a Digital Object Identifier (DOI) to help find and cite code" , failed itself to use the DOI when citing . Another reason, gathered from one of the authors by personal communication, but which I believe could be generalized, is that inclusion of the DOI (or other PID) was not part of the citation style of the publisher. In fact, it sometimes happens that publishers impose their own citation formats or standards, excluding the use of PIDs. Again, PIDs must be *used* and cited to persist. Citations help in achieving wide distribution, depending on good 'validatability', in order not to multiply errors and 'non-resolution' as a result.

Again, going back to the question of resolvability, the relationship between identifiers such as DOIs and URIs/IRIs is not always straightforward, and sometimes involves a chain of redirects ('303s'), before reaching a destination holding also the appropriate metadata.. Another reason resolvability may not be sufficient, even if the metadata is somehow in place, is that the file on the destination page resolved to is behind a paywall. In a case from December 2016, public domain content more than 110 years old was hidden behind a DOI-resolver charging 50$ for release of the content.

When someone in an ensuing Twitter conversation complained about this, an answering tweet seemed to mean, that was the price we have to pay for something as useful as DOIs. There was also a remark to the effect that every object gets one and only one DOI. But this is not necessarily true. It is possible to mint several DOIs for the same resource by different agents, such as Dataverse, Figshare, Zenodo etc. If not, that would give commercial publishers the opportunity to pro-actively seize whatever public domain content there is out there on the internet, quickly mint and assign their DOI to it and then lock it up behind paywalls. In any event, it would certainly not promote the use of PIDs instead of simple, ephemeral URLs for citation.

The DOI in question then was: [10.1080/00222930908692639](https://doi.org/10.1080/00222930908692639). The remedy proposed to "jump over" the paywall was by means of what was then *oadoi.org*, which still worked quite well. In several steps it eventually led, via an API, to an XML-file with a link to the freely accessible fulltext at <http://www.biodiversitylibrary.org/part/60220>. However, things have changed. When tried again in November 2018, the replacement *unpaywall.org* and oadoi-API for [10.1080/00222930908692639](https://api.oadoi.org/v2/10.1080/00222930908692639?email=j@s) was not working anymore for this DOI; the response received was: best\_oa\_location: null. But a free version of the resource sought, although no longer detected by *unpaywall.org*, could still be found, e.g., at [biodiversitylibrary.org](https://www.biodiversitylibrary.org/item/71907#page/71/mode/1up). However, often you will not find a free replacement copy of paywalled resources. An earlier "error message" (December, 2016) from *oadoi.org* for , said this holds for around 80% of scholarly articles. Maybe the situation has changed for the better since then. In January 2018, the unpaywall.org FAQ-pagestated: We find fulltext for 50-85% of articles, depending on their topic and year of publication. But, this message is no longer found there (in November 2018). And as we have seen in the example above, apparently in some cases *unpaywall.org* no longer finds open access fulltext alternatives, that were previously found by the *oadoi.org* application used then.

If *unpaywall.org* or an oadoi-API (e.g. <https://api.oadoi.org/10.3897/biss.2.25805?email=p@x> - the 'email' argument as given here may well be "fake", as long as it holds the @-character) fails to find a freely accessible publication (as with [https://api.oadoi.org/10.1007%2Fs10654-018-0449-x?email=n@y](https://api.oadoi.org/10.1007/s10654-018-0449-x?email=n@y)), an alternative might be to try the [identifiers.org SPARQL endpoint](http://identifiers.org/services/sparql) .But, it does not necessarily give us an open access URI in return. And it only works if the potential corresponding URIs have been assigned the property *owl:sameAs* just as the submitted subject URI. Unfortunately, in neither of our cases above are these conditions are met.

Assuming we have finally found a single seemingly reliable custodian for our PIDs and URIs, promising 24/7 resolution and top quality metadata, should we rest content with that? In law and journalism it is desirable not to judge by the testimony of only one witness or source. The evidence of at least two, mutually independent witnesses is generally preferred. *Multiple resolution* of any PID by several different proxy servers, as we already know, still means single parenthood, single custodianship of that lookup-table that has to be managed and updated in order for the PID to resolve as expected. Clark describes it as representing a stage in the evolution of PIDs, that will eventually be surpassed by a more mature age when we supply also *data types* to come with the PIDs, in order to make them more machine actionable . But we want more than that. We want backup custodians, metaphorical relatives or friends as caretakers. Providing multiple *access* to, or *identification* of resources through PIDs, that are capable of serving as trustworthy, competent, valid independent witnesses from different moments in time, at different sites, in different places is a good idea. Thus, we accept that an object may have multiple PIDs. Ideally these multiple PIDs should get to "know about" each other as a way towards interoperability.This can be achieved already, e.g. by means of linked open data (LOD), sameAs-relationships and tools provided by *n2t.net*,*unpaywall.org* and the *identifiers.org* SPARQL endpoint referred to above. Multiple identifiers from different namespaces for the same object may even be desirable in order to ensure interoperability in different environments.. It is also in line with the principle of the semantic web known as the *NUNA, Non-Unique Naming Assumption*, implying that things described in RDF data can have more than one name and any object may be identified by more than one URI, serving in RDF as 'names' of things.

However, this does not imply that any identifier, any PID is as good as the other. In fact, there are significant differences in quality between identifiers, particularly in terms of 'validatability' and 'meaningfulness'. We are getting there a bit later.

But first, having referred to linked data and sameAs-relationships as a possible solution to achieving interoperability, what about long-term sustainability? Are Linked Open Data (LOD), relying heavily on URIs, fit for survival? Archival records for long-term preservation need to be self-sustained, carrying meaning within themselves, while the references may no longer be resolvable. In e-archives compliant with the OAIS-model and Trustworthy Digital Repositories standards for self-sustenance, this means that URI strings lacking an inherently meaningful structure will often serve only as another set of dumb identifiers. Unless they can import some meaning from outside, through resolution or sameAs links, such opaque, non-resolvable URIs should henceforth rather be described as "non-semantic".

# **Which identifiers are FAIR enough?**

Just how persistent are PIDs really? Even if not always resolvable, are they in general still 'findable', well distributed over the internet in time and space? Are they 'validatable' (e.g. through fixed string-length, pattern-recognition, restricted character set, built-in checkdigit, built-in type?) Are they FAIR?

**Findability:** Beginning with the F for findability, for comparison we go back in time to 'old-fashioned' ISBNs, International Standard Book Numbers. Publicly declaring what type of objects they are meant to identify, ISBNs are rarely directly resolvable. But they are widely distributed, they have good *findability* in terms of *precision* hits, as seen by simple 'googling', with good survival rate, longer than the median age of web-pages 9.3 years. For example, look at ISBN 0-14-029161-X: *The Diversity of Life / Edward O. Wilson (2001)*. Simple googling of 014029161X, unprefixed and without hyphens results in 57/57 precision hits (date: 2017-01-30). ISBNs could also be searched in library catalogs, the most comprehensive of which is probably the [Karlsruhe Virtual Catalog – KVK worldwide](https://kvk.bibliothek.kit.edu/). Results of the query '014029161X', with the same unprefixed ISBN without hyphens yields 123/123 precision hits, recall being difficult to compute since in 55 of 72 catalogs the search could not be successfully processed or no records were found. To counteract the possibly unfair bias with a modern classic like this, we try instead an even older, and presumably less well-known example: ISBN: 2130381030. *L'Identité : séminaire interdisciplinaire dirigé par Claude Lévi-Strauss, 1974-1975* (Paris: PUF, 1983). Googling without prefix (2130381030) the precision is between 14/39 and 22/50; with prefix (ISBN2130381030) it reaches as high as 17/18 (date: 2017-01-30).

**Accessibility:** Data and (digital) objects are accessible only in so far as identifiers are findable or resolvable preferably to open access landing pages with either direct availability of resources, or sufficient metadata to direct the user to such an access point. In this respect DOIs are often, but not always, as good as or sometimes better than ISBNs (for obvious reasons regarding print only material), while gni-UUIDs as described above are all but useless.

**Interoperability** and **Re-usability** are both intimately associated with correctness, which can be helped by 'validatability', as argued above. We will look more into detail at the performance of different PIDs regarding this below.

**Archival Resource Key (ARK) Identifiers:** ARKs have a well defined syntax : [http://NMA/]ark:/NAAN/Name[Qualifier] , where *NMA* is a (changeable) *Name Mapping Authority*, a "host" or proxy resolving agent. This is not part of an ARK’s core identity, as marked out by the encompassing brackets and as shown by the example ARKs below resolving from two or more different NMAs (e.g., n2t.net and identifiers.org). The *NAAN* is the *Name Assigning Authority Number*, correseponding to the prefix starting with '10.dddd' in a DOI, and serving as a namespace for the following /Name. The NMA-supported *[Qualifier]* is not further defined in , but an example is given by the suffix *s3/f8.05v.tiff*, including also a file extension as we can see. As examples below, none of them having a qualifier, we find ARKs giving direct access to digital fulltext of Buffon’s *Histoire naturelle* at the BnF and a 20th Century Guide for mixing fancy drinks at the Internet Archive. In the third case, we see a resource with a prominent feature of ARKs, their possible *inflections*, here represented by the ‘?' at the end of the URI, giving both metadata for the 'document' itself and a photo of the Dallas Police Department from 1963; using ‘??’ gives “more” metadata, including a statement from the collection holding it. This inflection property of ARKs is intended to serve as a surrogate for the object, and could be a response to the requirement ensuing from the FAIR principle of accessibility (A2), that "metadata are accessible, even when the data are no longer available" , i.e., resources that have for some reason been removed from the net should at least leave a gravestone with metadata of last known whereabouts behind to the afterworld. Inflections also obviate the need for an ARK to point to a landing page (since metadata is available from the inflected form). Another interesting, perhaps unintentional feature of this particular case is that when the resolving agent, the NMA is changed from *texashistory.unt.edu* or n2t.net to either *identifiers.org*, the inflection appears to be ignored.

* [ark:/12148/bpt6k97497t](http://gallica.bnf.fr/ark%3A/12148/bpt6k97497t)
* [ark:/13960/t6c25cm5g](http://identifiers.org/ark%3A/13960/t6c25cm5g)
* [ark:/67531/metapth346793/?](https://texashistory.unt.edu/ark%3A/67531/metapth346793/)

What about the "validatability" of ARKs, then? As illustrated in the examples above, ARKs match the regular expression “^ark:\/[0-9]{5}\/\S+$”. After the “ark:/ddddd/”, there is no specific pattern or definite string-length of an ARK. The only restrictions on the Name and Qualifier parts "as strings of visible ASCII characters" is that they "should be less than 128 bytes in length" using "letters, digits, or any of these six [sic] characters: = # \* + @ \_ $", allowing also four more characters with reserved meaning: % - . /These general restrictions are not sufficient for efficient validation by means of a regular expression. While not required by the ARK specification, however, the first two examples above as well as most ARKs in the world (and some Handles), are generated using the NOID check digit algorithm [reference NOID manual page and proof], which was created in conjunction with ARKs. Despite NOID’s variable-length strings, they have stronger transcription error detection than ISBN or ISSN strings.

More heavily incumbent on an ARK seems to be the demand that, "[i]f longevity is the goal, it is important to keep the prefixes free of recognizable semantics". This, I believe, is a misconception. It is not the absence of semantic content that guarantees longevity; it is rather the continued *use* of the identifier that enhances its persistence, as observed again by above, and this use may well be promoted by at least some semantic content in the identifier, allowing a user to recognize it as an identifier of precisely that particular object that it is supposed to identify. The ARK specification discourages adding visible information to an ARK string, e.g., "[i]f longevity is the goal, it is important to keep the prefixes free of recognizable semantics". Unfortunately, this limits the use, and therefore persistence, of the identifier because such things as object type, date issued, and registrant are not meant to be included. The *Findability* by simple 'googling' and current *Accessibility* of the example ARKs above presently (Nov. 2018) still seems quite good. At least the first of these three examples seems to be well distributed, producing an impressive precision score of 25/25 by simple googling of '12148/bpt6k97497t' (each page in the hitlist actually containing a reference to the same document by Buffon in the Gallica collection). The second example apparently has a narrower distribution, but the few items found still display good precision, 4/4. The third example, without inflection, has been used extensively as a paradigmatic case, so should perhaps be considered outside competition here, but anyway also shows good precision. And, given the limited *validatability*, the degree to which the resources they identify will also be *Interoperable* and *Re-usable* will very much depend on the difficult to predict long-term sustainability of these two first FAIR properties of these ARKS.

**DOI:** DOIs can look just like anything. Here are some real cases, all at the time of writing resolvable and with multiple Findability also by simple googling, some of them pretty 'old', although they got their DOIs assigned fairly recently. One from 1977 (doi: 10.1177/030631277700700112) produces an impressive precision score of 68/68 (date: 2018-11-07), mostly due to it quite high citation rate, yielding hits for all the citing sources.

* [10.1007/978-3-319-07443-6\_39](https://doi.org/10.1007/978-3-319-07443-6_39)
* [10.1002/asi.23256](https://doi.org/10.1002/asi.23256)
* [10.1177/030631277700700112](https://doi.org/10.1177/030631277700700112)
* [10.1002/(SICI)1097-4571(199510)46:9<646::AID-ASI2>3.0.CO;2-1](https://doi.org/10.1002/%28SICI%291097-4571%28199510%2946%3A9%3C646%3A%3AAID-ASI2%3E3.0.CO;2-1)
* [10.1007/s11192-007-1682-3](https://doi.org/10.1007/s11192-007-1682-3)
* [10.1023/B:SCIE.0000018543.82441.f1](https://doi.org/10.1023/B%3ASCIE.0000018543.82441.f1)

Now, following are two DOIs from Wiley Online Library 1996 and Springer 2001 that still do not seem to resolve properly (neither on 2017-01-31, nor on 2018-11-11):

* [10.1002/(SICI)1520-6297(199601/02)12:1<67::AID-AGR6>3.3.CO;2-#](https://doi.org/10.1002/%28SICI%291520-6297%28199601/02%2912%3A1%3C67%3A%3AAID-AGR6%3E3.3.CO;2-)
* [10.1007/s00145-001-0001-](https://doi.org/10.1007/s00145-001-0001-x)**[x](https://doi.org/10.1007/s00145-001-0001-x)**

However, the two following DOIs, again from Wiley Online 1996 and 1998, that were similarly unresolvable at the same date (2017-01-31), are proof that with things like link-checking and maintenance, some PIDs might regain there resolvability later.

* [10.1002/(SICI)1520-6297(199601/02)12:1<67::AID-AGR6>3.3.CO;2-K](https://doi.org/10.1002/%28SICI%291520-6297%28199601/02%2912%3A1%3C105%3A%3AAID-AGR10%3E3.3.CO;2-K)
* [10.1002/(SICI)1520-6297(199811/12)14:6<475::AID-AGR5>3.3.CO;2-6](https://doi.org/10.1002/%28SICI%291520-6297%28199811/12%2914%3A6%3C475%3A%3AAID-AGR5%3E3.3.CO;2-6)

Obviously, all these DOIs, whether resolvable or not, vary substantially in string-length, from just 17 to over 60 characters, some involving abbreviations of journals or organisations, one an ISBN, and some containing characters in need of special XML-encoding, different from URI. Note that although the two last items in the first group are from the same journal, *Scientometrics*, they are quite different in structure. Anyway, all the above DOI examples are *valid* in accordance with the best we can offer as a regular expression restriction, with only partial pattern recognition: ^10\.[0-9]{4,}\/\S+$

meaning that any valid DOI must start by '10.' followed by a minimum of 4 digits, before the slash '/' and then a suffix of any length or characters, but no spaces in between.

But then, according to the same partial restriction, *defined by the regex above*, this entirely fake DOI is equally valid:

* 10.99999999/xxxxxxxx/x(y)x\:-{=?%%@@@@@

To be sure, there are other regular expression restrictions suggested for DOIs, those that are even more permissive (as DataCite 4.1, with the pattern value for doiType set to "10\..+/.+" , apart from not being PHP or JavaScript compliant, allowing also inline spaces, or the pattern registered for DOIs at *identifiers.org* as "^(doi\:)?\d{2}\.\d{4}.\*$" [MIR:00000019](https://www.ebi.ac.uk/miriam/main/collections/MIR%3A00000019), both of which also allow for the fake DOI above as valid, when tested in [regex101.com](https://regex101.com/)).There are other patterns that are more restrictive, but then obviously not catching all the now prevalent and permitted DOIs by one singular regular expression. , Thus, unlike ISBNs, DOIs are difficult to validate properly. Or rather, it is hard to find sufficiently discriminatory criteria to distinguish proper DOIs from fake ones. They have no fixed string-length, and few character set restrictions. All we can have is a partial pattern recognition; the more restrictive the validation rule or regular expression, the more it is likely to leave out extant DOIs.

**Handle:** The Handle identifier system, of which DOIs are only a special case, seems fairly easy and handy at first glance. Handles come in two different flavors. One is the semantically opaque, which has the structure: *Prefix/noid* (10079/sqv9sf1), where the NOID-part (for Nice Opaque Identifier ) is a short alphanumeric string from the restricted character set "0123456789bcdfghjkmnpqrstvwxz", with random minting order.The other flavor is the semantically transparent, which could be of three different types: the URL handle: *Prefix/local-PID* [(10079/bibid/123456)](https://hdl.handle.net/10079/bibid/123456) , the user handle: *Prefix/netid/netid*[(10079/netid/guoxinji)](https://hdl.handle.net/10079/netid/guoxinji), which as demonstrated here seems to be less persistent, as people tend to move, and the simpler group handle: *Prefix/group*[(10079/ISPS)](https://hdl.handle.net/10079/ISPS). While those of the second flavor might be more instantly "meaningful", providing context, how are Handles faring regarding Findability and Accessibility? The Findability by googling will again, as for other PIDs, largely depend on the use and citation rate of items, while the accessibility again rests largely on the maintenance of the lookup-table by the custodian. Even so, as just demonstrated, Handles may not always resolve to the page expected, especially when used as context dependent identifiers of individuals. In these cases, ORCiD ids should be preferred. What about the Interoperability and Re-usability of Handles then? Those of the NOID type, with a restricted character set, will in principle at least be effectively "validatable", to the extent that the "namespace" or minting agent restricts the string-length, as e.g. [2077/36687](https://hdl.handle.net/2077/36687) – Gothenburg university: 4/5 characters, and [10079/31zcrtn](https://hdl.handle.net/10079/31zcrtn) – Yale university: 5/7 characters. Those of the second, "semantic" flavor will apparently prove less "validatable" in the sense that there is no longer any fixed string-length or restricted character-set.

**UUID:** UUIDs v5 are used within the field of biodiversity taxonomy, as a complement to scientific names.They were introduced to the field in 2015 by the Global Names Architecture - GNA to convert scientific names to UUIDs. The arguments for using them instead of name strings for certain functions are that they save space as index keys in databases, they have a fixed string length (36 characters, including the dashes) while scientific names are of variable length. UUIDs do not suffer, as names sometimes do, from encoding problems that are difficult to detect, and they are more easily distinguishable one from the other than name strings for closely related species variants. Specifically, it is argued that “UUIDs v5 ... can be generated independently by anybody and a given name string always maps to the same UUID v5 string... Same ID can be generated in any popular language following well-defined algorithm.”[[20]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cphilipson-savesd2017revext.html#gna-2015)Note, however, that it is actually the specific *name* string that is identified here, not the object, the organism, or the 'thing itself'. Thus, the resulting UUID is completely dependent upon the particular scientific name string (with its encoding), it cannot be used as a bridge between different name forms for the same organism, telling us that they are naming the same object. This is due to the fact that it is generated by hashing a namespace identifier and name.As a result, UUIDs generated in this way by the gna name resolver, e.g. "707f84e1-e5b8-5063-8256-369ba9d72e13" for *Antiaris toxicaria* are next to useless as instruments of Findability, often yielding 0 hits by simple googling, all the while a search on the scientific *name* alone will give plenty of precision hits for the sought after organism, providing rich metadata for the 'thing' itself. Likewise, the same UUID is seldom or never Accessible, by being resolvable on its own. As an example, consider one of the most well studied organisms of all, the fruitfly *Drosophila melanogaster*. Using the Global Names Resolver to get a UUID v. 5 for *Drosophila melanogaster*, <gni-uuid>1bc2f359-47e4-5da6-a748-74676b7c8c5d</gni-uuid>, googling it either unprefixed or prefixed gives a zero result (0 recall, 0 precision, date: 2017-01-30). Trying instead the same UUID in a general search of all databases of NCBI, the US National Center for Biotechnology Information), we get 0 hits (2018-11-15): [1bc2f359-47e4-5da6-a748-74676b7c8c5d](https://www.ncbi.nlm.nih.gov/gquery/?term=1bc2f359-47e4-5da6-a748-74676b7c8c5d). Most notably, we get 0 hits in the [NCBI Taxonomy database](https://www.ncbi.nlm.nih.gov/taxonomy/?term=1bc2f359-47e4-5da6-a748-74676b7c8c5d), that on the face of it would seem to be the most relevant to our search.

By contrast, the UUID v5 is eminently "validatable", with a character set restricted to digits and lower case [a-f], and a fixed string length, 36 characters including hyphens, in a recognizable, precise pattern: "8-4-4-4-12", allowing for validation by a regular expression such as ^[a-f\d]{8}-[a-f\d]{4}-[a-f\d]{4}-[a-f\d]{4}-[a-f\d]{12}?$, or by means of an [online validator](http://www.freecodeformat.com/validate-uuid-guid.php). On the other hand, given the fact that these generated UUIDs are seldom or ever used for citation, and are not "fed back" to the source databases, where the corresponding scientific names were found, it is doubtful whether this "validatability" is also sufficient to make them qualify for *Interoperability* and *Re-usability.* To improve its findability and re-usability the UUIDs v5 issued by the Global Name Resolver could "ping-back" and assign themselves to the records in the biodiversity database sources they were drawn from and further use *schema.org* markup to get incoming links and a better ranking by search engines.

# **Why context?**

Generally speaking, although it is preferable that identifiers be findable and identifiable also in their unprefixed, pure form, typed identifiers give context by means of namespace prefixes of a metadata standard, a vocabulary or ontology. A typed identifier "introduces itself", telling us what kind of identifier it is, and what type of objects it is used for. Most importantly the namespace tells us what schema(s) and which rules should be used for its validation.

Page claimed that e.g. "dc:title" is adding "unnecessary complexity (why do we need to know that it's a "dc" title?)" in the JSON expression:

{ "@context": { "dc:title": "http://purl.org/dc/terms/title" },

 "dc:title": "Darwin Core: An Evolving Community-Developed Biodiversity Data

 Standard" }

A simple answer is that namespaces are important to retain meaning from context, serving as a key to interpretation for the future. Self-sustained long-term preservation should ideally mean in a case like this that the dc specification and schemas valid at the time be archived together with the records, [[32]](file:///C%3A%5CUsers%5Cjoph9849%5CDesktop%5Cphilipson-savesd17revext%5Cds-paper-558.html#li_sugimoto-2014) or at least that there is provenance metadata including timestamps and namespace of terms used. Metadatafiles in XML usually have a xsi:schemaLocation indicating which schema to validate against. This information, together with timestamped metadata elements such as 'dateIssued' should be sufficient to provide context. For JSON metadata there are name/value pairs such as { "protocol": "doi", ... "createTime": "2017-01-12T10:49:03Z", ...} that could fulfill the same function. And then, context is just as important for validation of records also in the present.

# **A "new" contextual, integrated, validatable PID?**

As seen in the case of Handle above, validatability sometimes comes at a cost: transparency lost. Are we forced to make a choice between the two? Can we create identifiers that are both fully validatable and at the same time more meaningful, providing context? So, here we suggest a *model* for a "new" PID, with a limited character set, at least for the object id part, defined by namespace specifications and schemas.

*Model*: [namespacePrefix].[objectType].[objectId: 10 chars].[issuedDate: YYYY-MM-DD].[registrant: ORCID or ROR-ID]

*Example (expression of this paper)*: fabio.PositionPaper.pp1255qv43.2018-11-12.0000-0001-5699-994X

It is a model of a contextual, validatable identifier, structured into ‘.’-separated modules (sub-strings). To make it easier to implement, and more generalizable, there are no character set or string-length restrictions for the first two modules, except that they should not contain the dot (.), which is the module separator. Nevertheless, this means already existing namespaces and object types could already be used to create a PID in accordance with this model.

The third module, the objectId (local ID) has a limited character set, selected to avoid ambiguous interpretations and, to avoid making local uniqueness case-dependent , restricted to lower case letter characters and digits. The full stop or dot (.) was chosen as module separator, since it works well in both xml- and http-environments, without encoding, and is not subject to confusion as sometimes hyphens and dashes (en-dash and em-dash) can be. It also works for tokenization of strings. The object type identified in the second module should belong to the initial namespace prefix. Every namespace can have as many object types as needed. Namespace schemas could also define valid *data types* for their different object types, thus supplying PIDs with *data types*, in order to make them even more machine actionable.

The scalability of this model will mainly depend on the 10-character objectId and the size of the permitted character set. An objectId limited to the proposed character set [a-kmnp-z0-9] will have 3410 permutations within each namespace (and possibly objectType), still better than e.g. a 7 character Handle with NOID.

The objectId module, thus, could be validated separately by a regular expression restricted to ^[a-kmnp-z0-9]{10}$. It may also be part of a more comprehensive validation schema, involving a random or pattern based minting algorithm, preferably including a checkdigit, with different rules invoked for different namespace contexts, checking also for example the correspondence between namespace (module 1) and objectType (module 2) as in this still crude Schematron schema:

<schema xmlns="http://purl.oclc.org/dsdl/schematron" queryBinding="xslt2">

 <ns prefix="rdf" uri="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>

 <ns prefix="fabio" uri="https://w3id.org/spar/fabio"/>

 <ns prefix="local" uri="local"/>

 <pattern>

 <rule id="newPid-rule" context="local:newPid">

 <let name="objectType" value="for $i in (.) return tokenize($i,'\.')[2]"/>

 <let name="objectId" value="for $i in (.) return tokenize($i,'\.')[3]"/>

 <let name="x" value="'https://w3id.org/spar/fabio'"/>

 <let name="objectTypeList" value="(for $i in (doc($x)//rdf:type[@rdf:resource='http://www.w3.org/2002/07/owl#Class']/parent::rdf:Description/@rdf:about)

 return substring-after($i,'fabio/'))"/>

 <let name="objectTypeString" value="string-join($objectTypeList,',')"/>

 <assert test="matches($objectId,'^[a-kmnp-z0-9]{10}$')">

 An identifier of type 'newPid' must have as its third module a namespace unique objectId of 10 characters from the set [a-kmnp-z0-9].</assert>

 <assert test="matches($objectTypeString,$objectType)">The objectType, the second module of the newPid must belong to the namespace of the first module.

 </assert>

 </rule>

 </pattern>

<schema>

One might consider generalizing such a validation schema to the extent possible, so that the namespace URI in $objectTypeList, from which the $objectType should be drawn, was automatically construed based on the namespacePrefix (module 1) of the newPid instance to be validated. This could be achieved by having the namespacePrefix expressed as a link with a namespace URI, e.g. such as [fabio](https://w3id.org/spar/fabio.xml) in our case above. But that would also make the validation schema a bit more complicated, notably in the tokenization and separation of modules.

It is also conceivable, in order to allow for integration of already existing identifier schemes, that a namespace sets its own character set and string-length restrictions, to be declared by the validation rules of that namespace. For "narrow" namespaces, lacking defined diverse object types, possibly since they comprise basically only one type of object (as for ISBNs and ISSNs) we suggest as a default second module value *'NOT'* = No Object Type. So we could have an IGSN, *International Geo Sample Number* , with string-length of objectID set to 9, expressed in this model:

*Example*: IGSN.NOT.IECUR0002.2005-03-31.gswa-library

The identifier should be fully validatable, as a whole or in parts (modules), in the corresponding namespace(s). The last two modules are optional, but they are meant to offer built in data provenance. For organisation identifiers, we hope that the recently launched ROR-IDs will become a global standard, like the ORCID for persons. Then we could replace the last module in the IGSN-PID above with "05h2dda38".

The resulting PIDs should be minted within the corresponding namespaces, which would also be the 'custodians' and resolving authorities of their respective PIDs, responsible for uniqueness within the namespace. Another task would be to monitor and assign sameAs-properties to PIDs that refer to the same 'thing' in other namespaces.

It has been suggested that in order to build more connected, cross-linked and digitally accessible Internet content it is necessary to assign recognizable, persistent, globally unique, stable identifiers to ... data objects.. The model proposed here aims to make "new" PID strings that are universally unique, and stable, but with enough meaning to be fully recognizable in a well-known context, and with a good potential for backup.

# **Conclusions**

The purpose of this paper was to analyse some of the more prevalent general PIDs used in scholarly communication, identify some of their shortcomings and find out how PIDs could be made more FAIR. Real examples of PIDs were analysed to find out what additional requirements there might be to make them fully Findable, Accessible, Interoperable and Re-usable - FAIR. The "novelty" of the paper, if any, is then the "widening" of the FAIR principles to include also Findability as rate of distribution or dissemination (e.g. as measured by means of 'googling') and Interoperability or Re-usability to include also 'validatability'. Further, as against earlier insistence on the opaqueness of PIDs as a warrant for persistence, we argued for the importance of adding enough meaning to PIDs, through namespace prefixes and object types, so as to enhance their future use, distribution, findability and interpretability, and to safeguard against failed resolvability. The custodianship and minting of PIDs, we suggested to be the responsibility of namespaces, as these are already assuming the administration of specifications, validation schemas, vocabularies or ontologies, and should be well qualified for the task. The minting algorithm, the patterns for PID-recognition, restriction in character set, string-length (with possible checkdigit) of objectId module should all be part of the validation schemas. These namespaces should then be able to register their schemes with *n2t.net* or *identifiers.org*, as already happens. And there might be several services such as the SPARQL endpoint of *identifiers.org* for registering sameAs-links. To create, maintain and make our PIDs truly persistent, widely used and FAIR should be a cooperative effort of the whole scholarly community.

# **References**

1. California Digital Library (2018). *Archival Resource Key (ARK) Identifiers.* <http://n2t.net/e/ark_ids.html>
2. Kunze, J. & Rodgers, R. (2008). *The ARK Identifier scheme.* [http://n2t.net/e/arkspec.txt](https://n2t.net/ark%3A/13030/c7cv4br18)
3. Clark, J. (2016). *PIDvasive:\_What's possible when everything has a persistent identifier?* PIDapalooza, November 10, 2016. Retrieved Jan 16, 2017.<https://doi.org/10.6084/m9.figshare.4233839.v1>
4. Catalogue of Life: Annual Checklist(2015). *Asterolibertia gibbosa (Gaillard) Hansf. 1949.* <http://www.catalogueoflife.org/annual-checklist/2015/details/species/id/4f5bf9e96f36e1c530b147c7105e865b>
5. Coyle, K. et al.(2014). *How Semantic Web differs from traditional data processing.* RDF Validation in the Cultural Heritage Community. International Conference on Dublin Core and Metadata Applications, Austin, Oct. 2014. Date accessed: 24 Mar. 2017. [http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/311](http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/311/)
6. Cruz, M., Kurapati, S., & Turkyilmaz-van der Velden, Y. (2018). *The Role of Data Stewardship in Software Sustainability and Reproducibility.*Zenodo. 2018-09-14. [https://doi.org/10.5281/zenodo.1419085](http://doi.org/10.5281/zenodo.1419085)
7. DataCite Metadata Working Group. (2017). *DataCite Metadata Schema 4.1*. <https://doi.org/10.5438/0015>
8. Doorn, P., Dillo, I. (2017). *Assessing the FAIRness of Datasets in Trustworthy Digital Repositories: A Proposal.* IDCC Edinburgh, 22 February 2017. <http://www.dcc.ac.uk/webfm_send/2481>
9. Duerr, R.E. et al. (2011). (2011). *On the utility of identification schemes for digital earth science data: an assessment and recommendations .* Earth Science Informatics 4:139. ISSN: 1865-0473 (Print) 1865-0481 (Online) <https://doi.org/10.1007/s12145-011-0083-6>
10. Dunning, A., de Smaele, M., Böhmer, J. (2017). *Are the FAIR Data Principles fair?* Practice Paper. 12th International Digital Curation Conference (IDCC 2017), Edinburgh, Scotland, 20 - 23 February 2017. <https://doi.org/10.5281/zenodo.321423>
11. Fenner, M. (2016). *Cool DOI's*. DataCite Blog. <https://doi.org/10.5438/55e5-t5c0>
12. Force11 (2016a). *The FAIR Data Principles.* <https://www.force11.org/group/fairgroup/fairprinciples>
13. Force11 (2016b). *Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version B1.0.*<https://www.force11.org/fairprinciples>
14. FAIRMetrics (2018). *FM-F2* <https://purl.org/fair-metrics/FM_F2>
15. FAIRMetrics (2018). *FM\_R1-3*. <https://purl.org/fair-metrics/FM_R1.3>
16. Data Citation Synthesis Group, Martone M. (ed.)(2014). *Joint Declaration of Data Citation Principles* San Diego, CA: FORCE11 <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
17. Force11 (2016). *Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0* San Diego, CA: FORCE11 <https://www.force11.org/node/6062/#Annex6-9>
18. Gertler, A., Bullock, J. (2017). *Reference Rot: An Emerging Threat to Transparency in Political Science.* The Profession. [http://doi.org/10.1017/S1049096516002353](https://doi.org/10.1017/S1049096516002353)
19. Gilmartin, A. (2015). *DOIs and matching regular expressions*. Crossref Blog, 2015-08-11. <https://www.crossref.org/blog/dois-and-matching-regular-expressions/>
20. Global Names Architecture - GNA (2015). *New UUID v5 Generation Tool -- gn\_uuid v0.5.0.* <http://globalnames.org/news/2015/05/31/gn-uuid-0-5-0/>
21. Global Names Architecture - GNA (2015b). *Global Names Resolver* <http://resolver.globalnames.org/>
22. Guo, Xinjiang (2016). *Yale Persistent Linking Service* PIDapalooza, November 10, 2016. Retrieved Jan 16, 2017. <https://doi.org/10.6084/m9.figshare.4235822.v1>
23. Guralnick, R. et al. (2015). *Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data.* ZooKeys 494: 133–154. <https://doi.org/10.3897/zookeys.494.9352>
24. Haak, L. et al. (2012). *ORCID: a system to uniquely identify researchers.* Learned Publishing, 25: 259–264. <https://doi.org/10.1087/20120404>
25. Hayes, C. (2016). *oaDOI: A New Tool for Discovering OA Content*. Scholars Cooperative, Wayne State University. <http://blogs.wayne.edu/scholarscoop/2016/10/25/oadoi-a-new-tool-for-discovering-oa-content/>
26. Hayes, C. (2017). *Unpaywall: A New OA Discovery Tool*. Scholars Cooperative, Wayne State University. <https://blogs.wayne.edu/scholarscoop/2017/03/20/unpaywall/>
27. Hennessey, J., Xijin Ge, S. (2013). *A Cross Disciplinary Study of Link Decay and the Effectiveness of Mitigation Techniques.* Proceedings of the Tenth Annual MCBIOS Conference. BMC Bioinformatics, 14(Suppl 14):S5. <https://doi.org/10.1186/1471-2105-14-S14-S5>
28. Jones, SM., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., Grover, C. (2016). *Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content.* PLoSONE 11(12): e0167475. [https://doi.org/10.1371/journal.pone.016747](https://doi.org/10.1371/journal.pone.0167475)
29. Kille, L.W. (2015). *The growing problem of Internet "link rot" and best practices for media and online publishers.* <https://journalistsresource.org/studies/society/internet/website-linking-best-practices-media-online-publishers>
30. Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva L., Zhou, K., Tobin, R. (2014). *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.* PLoS ONE 9(12): e115253. <https://doi.org/10.1371/journal.pone.0115253>
31. Kunze, J., Russell, M. (2006). *Noid - search.cpan.org.* <http://search.cpan.org/~jak/Noid/noid>
32. Li, C.& Sugimoto, S. (2014). *Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata.* Proceedings of the International Conference on Dublin Core and Metadata Applications (Austin TX, 2014). <http://dcpapers.dublincore.org/pubs/article/view/3709>
33. McMurry, JA. et al. (2017). *Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data.* PLoSBiol 15(6):e2001414.. [http://doi.org/10.1371/journal.pbio.2001414](https://doi.org/10.1371/journal.pbio.2001414)
34. Page, R. (2016). *Towards a biodiversity knowledge graph.* Research Ideas and Outcomes 2: e8767 (07 Apr 2016). [http://doi.org/10.3897/rio.2.e8767](https://doi.org/10.3897/rio.2.e8767)
35. Paskin, N. (1999). *Toward Unique Identiﬁers.* Proceedings of the IEEE 87(7):1208 - 1227. <https://doi.org/10.1109/5.771073>
36. Patterson, D. et al. (2016). *Challenges with using names to link digital biodiversity information.* Biodiversity Data Journal 4: e8080 (25 May 2016). <https://doi.org/10.3897/BDJ.4.e8080>
37. Philipson, J. (2017). *About a BUOI: joint custody of persistent universally unique identifiers on the web, or, making PIDs more FAIR.* SAVE-SD 2017 <http://cs.unibo.it/save-sd/2017/papers/html/philipson-savesd2017.html>
38. Philipson, J. (2019). *The Red Queen in the Repository: metadata quality in an ever-changing environment.*IDCC 2019. (In press). <https://doi.org/10.5281/zenodo.2276777>
39. ROR (2019). *ROR* <https://ror.org/about/>
40. SESAR - System for Earth Sample Registration (2017). *What is the IGSN?* <http://www.geosamples.org/aboutigsn>
41. Unpaywall.org (2018). *Frequently Asked Questions* <http://unpaywall.org/faq>
42. Wikipedia (2017a). *Link rot.* (last modified on 13 March 2017, at 17:46. Retrieved 2017-03-14.) <https://en.wikipedia.org/wiki/Link_rot>
43. Wikipedia (2017b). *Universally unique identifier.* (last modified on 29 January 2017, at 15:28. Retrieved 2017-01-30.) <https://en.wikipedia.org/wiki/Universally_unique_identifier>
44. Van de Sompel, H., Klein, M., Jones, S.M. (2016). *Persistent URIs Must Be Used To Be Persistent.* WWW 2016. [arXiv:1602.09102v1 [cs.DL] 29 Feb 2016](https://arxiv.org/abs/1602.09102v1)
45. Van de Sompel, H. (2018).*cite-as: A Link Relation to Convey a Preferred URI for Referencing* <https://datatracker.ietf.org/doc/draft-vandesompel-citeas/>
46. Wass, J. (2016). *When PIDs aren't there. Tales from Crossref Event Data.* PIDapalooza, Reykjavik, November 2016. Retrieved: 11:57, Mar 20, 2017 (GMT). <https://doi.org/10.6084/m9.figshare.4220580.v1>
47. Wass, J. (2017). *URLs and DOIs: a complicated relationship.* CrossRef Blog, 2017 January 31. <https://www.crossref.org/blog/urls-and-dois-a-complicated-relationship/>
48. Wilkinson, M. et al. (2016). *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data 3:160018. [http://doi.org/10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)
49. Wilkinson, M., Schultes, E., Bonino, L., Sansone, S., Doorn, P. & Dumontier, M. (2018, July 4). *FAIRMetrics/Metrics: FAIR Metrics, Evaluation results, and initial release of automated evaluator code.* Scientific Data. Zenodo. [http://doi.org/zenodo.1305060](http://doi.org/10.5281/zenodo.1305060)
50. Wimalaratne, S. et al. (2015). *SPARQL-enabled identifier conversion with Identifiers.org* Bioinformatics, 31(11), 2015, 1875–1877. [http://doi.org/10.1093/bioinformatics/btv064](https://doi.org/10.1093/bioinformatics/btv064)
51. Zhou, K. et al. (2015). *No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving.* In: Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries. JCDL '15, p. 233-236. [http://doi.org/10.1145/2756406.2756940](https://doi.org/10.1145/2756406.2756940)