Plomp, E. 2020. Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments. *Data Science Journal,* 19: 46, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2020-046

ESSAY

Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments

Esther Plomp
Delft University of Technology, Faculty of Applied Sciences, NL
e.plomp@tudelft.nl

The uptake of Persistent Identifiers (PIDs) has increased in recent years and has improved the Findability, Accessibility, Interoperability and Reusability (FAIR) of various research related objects (e.g., data, software, researchers and research organisations). The uptake of PIDs for physical aspects of research (such as samples, artefacts, reagents and analyses instruments) has thus far been embraced primarily for use in the fields of Earth and life Sciences. Wider adoption of PIDs for physical aspects of research can improve the findability and accessibility of these resources, which will allow for data to be put into more detailed context. By using PIDs all the information about a sample or artefact could be more easily available in a single location, allowing for persistent links to other sources of relevant information. Through the use of interoperable (metadata) standards and shared forms of documentation it will be easier to collaborate across multiple disciplines and the reusability of resulting data and the physical samples and artefacts themselves will improve. Wider adoption of PIDs for physical aspects of research is challenging, as research communities will have to work together to establish relevant standards that are meaningful across multiple domains. The infrastructure for wider adoption already exists, it is now up to research communities to adopt standards and PIDs for the physical aspects of their research and up to funding and research institutes to support this broader adoption.

Keywords: FAIR; physical samples; persistent identifiers; data; PIDs

1. Introduction

Recent years have seen an increased focus on improved research data management in scholarship. It no longer suffices that the data supporting the conclusions of publications are 'available upon request'. Data being available upon request often results in data loss as the data is not always readily available or the authors can no longer be contacted (Vines et al, 2014). Providing these data as supplementary materials makes the data more accessible than having data available on request, however, long term preservation of supplementary materials is challenging. Supplementary materials are not consistently accessible and have limited standardisation in the file formats used and in their internal organisation, hampering systematic review or tracking of the data (Anderson et al, 2006; Evangelou et al, 2005; Santos et al, 2005). The predominant PDF file format of supplementary files is not always ideal for reuse, as data collection and analysis generally takes place in other file formats (Kwon, 2019). Furthermore, supplementary materials are not always persistently available, leading to broken links and possible data loss (Kwon, 2019).

Instead, all data supporting scientific results are increasingly made available through data repositories (Digital Science et al, 2019; European Commission, 2017; Stall et al, 2019). Making the data available by archiving it in a data repository follows the recommendation of the FAIR principles (Wilkinson et al, 2016). The FAIR principles provide guidance on how to make data Findable, Accessible, Interoperable and Reusable. These principles recommend that data should be 'Findable' on the internet, using a persistent identifier (PID) that allows citation and tracking of the data. The information about the data (metadata) should be 'Accessible'. The data should be in commonly used and preferably open file formats and described in standardised vocabularies to be 'Interoperable' with other data. By accompanying the data with proper

documentation and a user license the data can become 'Reusable' for other researchers, facilitating collaboration and maximising impact of the research outputs.

The FAIR principles "apply not only to 'data' in the conventional sense, but also to the algorithms, tools and workflows that led to that data" (Wilkinson et al, 2016). In the Beijing Declaration the term data is used "very broadly, to comprise data (stricto sensu) and the ecosystem of digital things that relate to data, including metadata, software and algorithms, as well as physical samples and analogue artefacts…" (Hodson et al, 2019). In the European Code of Conduct for Research Integrity, research data is also generally described as "research materials in all their forms (encompassing qualitative and quantitative data, protocols, processes, other research artefacts and associated metadata) that are necessary for reproducibility, traceability and accountability" (ALLEA, 2017). The Beijing declaration and European Code of Conduct refer to physical samples and artefacts as data, which can include samples such as biological specimens of plants and insects, minerals, soil, sediments, rocks, water, air, art, maps and physical texts, archaeological and synthetic materials, and tissues from humans and animals. The application of the FAIR principles to these physical samples, as well as reagents and instrument data, can address several challenges that these disciplines are currently experiencing. For example, it can be challenging to find information about samples that have been used in previous research. Often, naming conventions of physical samples and artefacts are not formalised and as a result, sample names can be ambiguous and heavily reliant on personal preferences, as well as subject to name changes over the course of a sample's lifecycle. This means that it is possible that different samples can be assigned the same name, or that a single sample has multiple names that are difficult to relate to each other which makes it difficult to track samples or resources across studies (Bandrowski et al, 2015; Devaraju et al, 2017; Hsu et al, 2020). Even if samples and artefacts follow a formalised naming convention, the detailed sample or resource descriptions are often not publicly available as they are rarely listed in the published literature (Bandrowski et al, 2015; Hills, 2015). Furthermore, catalogues or databases that offer this detailed information are usually not (publicly) available (Devaraju et al, 2017). This results in a loss of information that hampers the interpretation and reuse of research that is based on physical resources.

There thus remains a need to extend the FAIR principles to physical samples, artefacts, reagents, and analysis instruments, as these are essential for several research domains. Information generated using physical samples should be well documented and persistently available so that others are able to find the data, as well as verify and reuse the data. Ensuring that these samples and artefacts are available for wider reuse is more efficient as new sample and field campaigns are costly in terms of time and resources and not always possible. Collection and curation of samples and artefacts is a time-consuming endeavour and deserves recognition in the form of attribution, which could be enabled by making physical resources citable. Making information and data generated from physical samples more widely available in a standardised manner will facilitate collaboration across different research groups and disciplines, as it will be easier to identify which analyses have already been performed and see where the gaps in knowledge persist.

2. Findable and Accessible: Persistent identifiers

In order to expand the FAIR principles to physical aspects, physical data should be Findable. This can be done through the use of persistent identifiers (PIDs), long-lasting references on the internet to files, web pages, or other objects. These references will remain functional, ensuring access to the digital object. PIDs, such as Digital Object Identifiers (DOIs), have been in place for twenty years now and have been primarily used for published manuscripts.[1] Recent years have seen the introduction for PIDs for multiple research components, e.g., data and software, but also the more physical aspects of research such as research activities, research and funding organisations, as well as researchers themselves (**Table 1**).

All these persistent identifiers can be linked to each other, just as they are all related in the research life cycle. Project FREYA[2] has been working on a PID Graph that allows standardised cross linking between publications, data, researchers, institutes and funders (Fenner and Aryani, 2019). This work could be extended to include links to the instruments, reagents, artefacts and physical samples that are analysed, in order to be more representative of the entire research life cycle and to be able to integrate all the information of the data ecosystem (**Figure 1**).

[1] https://www.doi.org/doi_handbook/1_Introduction.html.
[2] https://www.project-freya.eu/en.

Table 1: Overview of persistent identifiers for physical research components, such as samples, resources, instruments, funding and research institutes and researchers.

Persistent identifier	Full name	Object	Webpage	Starting year
IGSN	International Geo Sample Number	Physical samples	https://www.igsn.org/	2007
ORCID	Open Researcher and Contributor IDentifier	Researchers	https://orcid.org/	2009
RRID	Research Resource Identifiers	Resources (antibodies, model organisms and software projects)	https://www.rrids.org/	2014
FundRef	Open Funder Registry	Funder	https://gitlab.com/crossref/open_funder_registry	2016
RAiD	Research Activity Identifier	Research activities	https://www.raid.org.au/	2017
RoR	Research Organization Registry	Research organisations	https://ror.org/	2019
PIDINST	Instruments	Persistent Identification of Instruments	https://www.rd-alliance.org/groups/persistent-identification-instruments-wg	2020

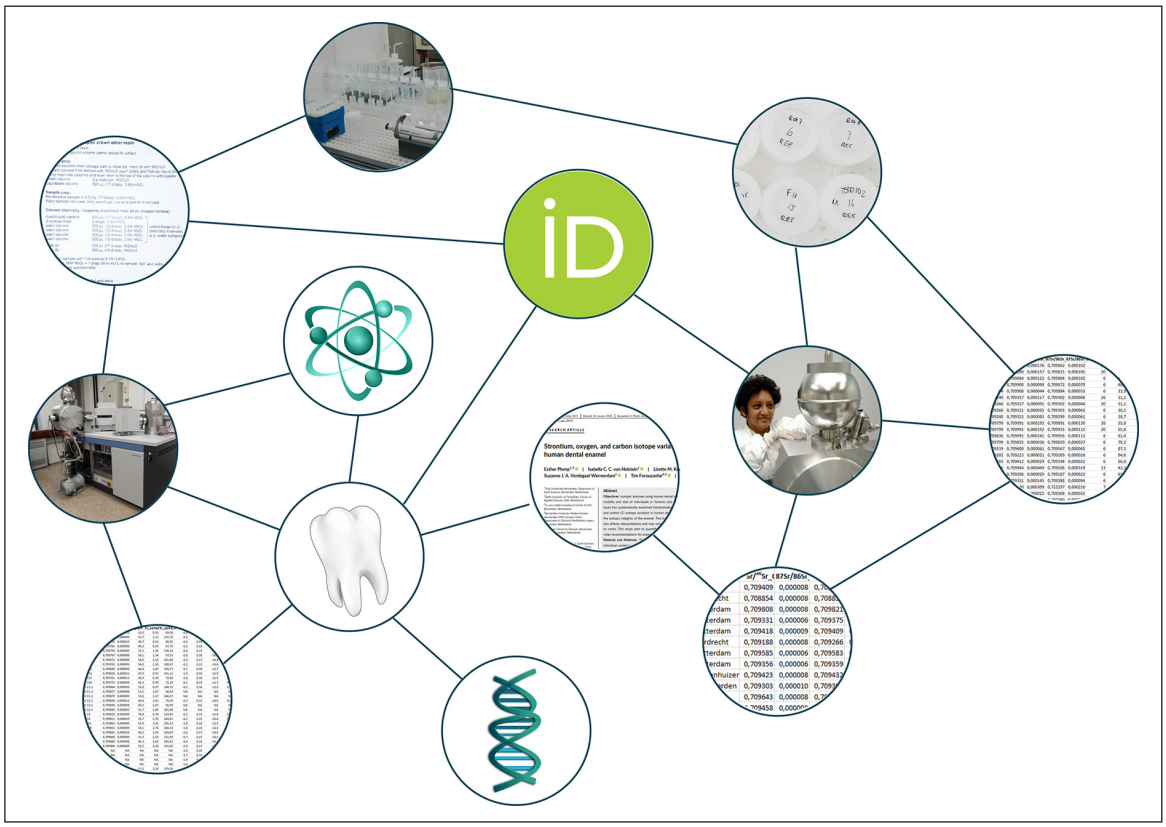

Figure 1: Linking different aspects of research, including physical samples, artefacts, instruments and reagents (after the PID Graph (Fenner and Aryani, 2019)). Photo credit: Esther Plomp/Dean Calma – IAEA[3].

Persistent identifiers for physical aspects of research

Persistent identifiers for physical aspects of research are already available or promoted by several initiatives. Some examples are listed (in chronological order) below:

[3] https://www.flickr.com/photos/iaea_imagebank/8160673319. (CC BY-SA 2.0).

The **International Geo Sample Number (IGSN)** makes samples discoverable and accessible since 2007, primarily for the Earth Sciences. The implementation of IGSN ensures that measurements based on samples can be repeated and validated, or that new measurements can be made (Devaraju et al, 2017). The initiative was established with funding from the National Science Foundation and uses the handle system (based on the DataCite metadata scheme) to assign persistent identifiers to physical samples, sampling features from which the sample was taken, a collection of samples and subsamples (samples derived from an existing sample) (Devaraju et al, 2017; Lehnert et al, 2019). The landing page of the persistent identifiers contains more detailed information of the registered resources (Devaraju et al, 2017).

Research Resource Identifiers (RRIDs) were introduced in 2014 and provide information on research resources (reagents, materials and tools used to produce the findings of the study) used in biomedical literature (Bandrowski et al, 2015). The successful uptake of RRID, now used in over 120 journals, is mainly due to the journals introducing requirements and instructions to authors (Bandrowski et al, 2015; Hsu et al, 2020). Thanks to the use of RRIDs, resources used can now be identified in 95% of the cases, compared to 50% without RRIDs (Hsu et al, 2020). To make the curation of these resources more manageable, a semi-automated curation tool was developed to validate RRIDs in published papers: SciBot (Babic et al, 2019; Hsu et al, 2020).

The foundations of the **Distributed System of Scientific Collections (DiSSCo)**[4] were laid in 2015. DiSSCo is an European effort to make biodiversity data more FAIR through the use of persistent identifiers (Digital Collection and Digital Specimen objects).

Persistent Identification of Instruments (PIDINST) aims to set up PIDs for operational scientific instruments to provide analysis metadata that helps to set the data into context (Stocker et al, 2020). Efforts have been undertaken by the PIDINST working group members of the Research Data Alliance since 2017, with primarily Earth Science use cases (Stocker et al, 2020). PIDINST provides metadata such as the instrument's name, a textual description, the institution where the instrument is situated, the manufacturer and other entities/objects that relate to the instrument (Stocker et al, 2020). The PIDINST schema facilitates links among instruments, journal articles, datasets and other research objects (Stocker et al, 2020).

3. Interoperable: Envisioned challenges

The infrastructure is available to implement persistent identifiers for physical aspects of research, but broader adoption beyond the Earth and Life Sciences (Bandrowski et al, 2015; Devaraju et al, 2017; Hsu et al, 2020; Stocker et al, 2020) is currently still limited.

Adopting persistent identifiers in different research fields will be challenging, as each discipline has its own data culture and jargon that complicates the use of shared schemas, registries, controlled vocabularies and ontologies (Poirier and Costelloe-Kuehn, 2019). There will be no common standard that is meaningful for the variety of experimental techniques used across different subdisciplines (Stocker et al, 2020). Developments on exploring a core requirement set of metadata, or a 'bullseye' that defines a common core kernel (Wyborn et al. 2020), that can be extended with discipline specific community standards are needed for a broader adaptation of persistent identifiers for physical samples, artefacts, reagents and instruments (**Figure 2**). Without a certain degree of standardisation it will be difficult to provide interoperability and compare samples, reagents and instruments across disciplinary boundaries. Efforts are being undertaken by IGSN (Sloan Foundation IGSN 2040 Project) to support disciplines other than the Earth Science community in using persistent identifiers for physical samples (Aronsohn, 2018). This will require adaptation of the current IGSN policies and metadata/categorisation schemes to support a broader diversity of sample times from these disciplines (Aronsohn, 2018; Lehnert et al, 2019). A solution to address the diversity of multiple disciplines would be to have more detailed information (such as information on sub- or composite samples, field programmes, protocols, technical set ups of instruments), next to the core metadata requirements, available on a landing page associated with the persistent identifier. This ensures that the information needed by discipline specific research communities is also persistently available (Stocker et al, 2020).

Research communities will need to establish their requirements for documentation of physical aspects of research and establish (meta) data standards where needed. Professional societies are in a good position to start or to facilitate this work, but grass root developments should also be encouraged through funding opportunities, such as the support from the Sloan Foundation for IGSN (Aronsohn, 2018). Where

[4] https://www.dissco.eu/.

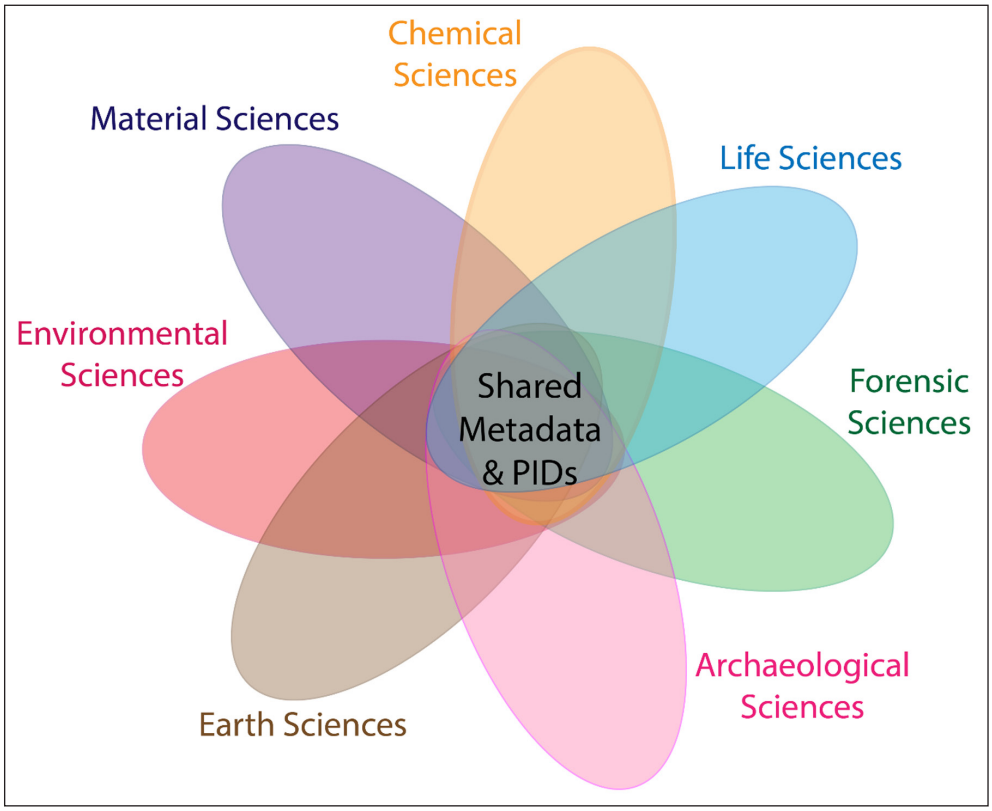

Figure 2: To ensure interoperability across disciplinary boundaries (e.g., Archaeological Sciences, Material Sciences, Earth Sciences, Environmental Sciences, Life Sciences, Forensic Sciences, Chemical Sciences), it is important to have a degree of shared metadata and PIDs. Next to the shared metadata, discipline specific information and documentation can be made available.

disciplinary boundaries need to be crossed, the Research Data Alliance[5] and FORCE11 provide important international and inter-disciplinary places for (meta)data discussions.

4. Reusable: Community engagement

For the successful adoption of persistent identifiers for physical aspects of research, increasing awareness about the available infrastructure and established standards is needed. This includes highlighting the benefits that persistent identifiers bring to researchers. Examples are numerable, as persistent identifiers and associated landing pages improve the sharing of contextual data, which can be interpreted across disciplines. These records are citable, providing attribution and visibility for sharing this type of data or providing access to physical samples and artefacts. Another example of a direct incentive to adopt PIDs are requirements from publishers to use them in publications. As the introduction of RRIDs highlighted (Bandrowski et al, 2015; Hsu et al, 2020), it is important that researchers are provided with training or instruction templates to facilitate implementation and registration of persistent identifiers. Herein lies an opportunity for publishers, institutional research data management or repository support staff.

The efforts that are involved in making physical data FAIR should be rewarded. This can be done by taking FAIR physical data in to account in promotion and tenure processes at institutes and in the assessment of research proposals. This would require the inclusion of physical samples and research data in the definition of data in policies of funding and research institutes, following the example of the Beijing Declaration. Similarly, (inter)national integrity codes can follow the definition of data from the European Code of Conduct for Research Integrity. The inclusion of physical data in policies recognise the value of physical samples and resources and awareness could be further increased by including physical aspects in data management

[5] See the work by the Physical Samples and Collections in the Research Data Ecosystem Interest Group (https://www.rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig) and the Persistent Identification of Instruments Working Group (https://www.rd-alliance.org/groups/persistent-identification-instruments-wg).

plans. Funders can furthermore ensure the long term sustainability of the infrastructures that enable FAIR physical data by offering financial support through dedicated funds or calls.

To further improve the reusability of physical samples and artefacts, their accessibility to other research groups could be improved. There are several institutes that already provide external access to their physical samples. For example, biobanks make samples and data available for reuse in medical research. The NASA lunar sample building houses and prepares the Apollo samples for shipment to researchers, with nearly 400 samples distributed per year to research and teaching projects.[6] An important role could be played by museums and institutional repositories in the curation of physical samples and artefacts, ensuring preservation and access not only to the collected data but also the physical resources themselves. Initiatives such as the Global Sustainability Coalition for Open Science Services (SCOSS)[7] could potentially improve the sustainability of such infrastructures.

5. Conclusion

The application of the FAIR principles to physical aspects of research can greatly improve the preservation, interpretation and reusability of this type of research. To facilitate the adoption of persistent identifiers for physical aspects of research, research communities will need to establish leading practices and (meta) data standards for the collection of information. Researchers can be supported in these efforts by funders, publishers, institutional support staff, and professional organisations. Existing initiatives in Earth and Life Sciences provide important examples of how persistent identifiers will allow others to verify and reuse the data, as well as the physical samples and artefacts themselves. Implementation of persistent identifiers and standardisation in documentation will result in greater impact of research that involves physical resources, as well as increased visibility for researchers that make this type of data available.

Competing Interests

The author has no competing interests to declare.

References

ALLEA. 2017. The European Code of Conduct for Research Integrity, Revised Edition. Berlin: ALLEA – All European Academies. Available at https://allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf [Last accessed 29 June 2020].

Anderson, NR, Tarczy-Hornoch, P and **Bumgarner, RE.** 2006. On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics*, 7(1). DOI: https://doi.org/10.1186/1471-2105-7-260

Aronsohn, MD. 2018. Sloan Foundation Grant will help support open and transparent science. Available at https://web.archive.org/web/20191227102217/https://blogs.ei.columbia.edu/2018/07/20/sloan-foundation-grant-open-science/.

Babic, Z, Capes-Davis, A, Martone, ME, Bairoch, A, Ozyurt, IB, Gillespie, TH and **Bandrowski, AE.** 2019. Incidences of problematic cell lines are lower in papers that use RRIDs to identify cell lines. *eLife*, 8. DOI: https://doi.org/10.7554/eLife.41676

Bandrowski, A, Brush, M, Grethe, JS, Haendel, MA, Kennedy, DN, Hill, S, Hof, PR, Martone, ME, Pols, M, Tan, S, Washington, N, Zudilova-Seinstra, E, Vasilevsky, N and **Resource Identification Initiative Members are listed here: https://www.force11.org/node/4463/members.** 2015. The Resource Identification Initiative: A cultural shift in publishing. *F1000Research*, 4: 134. DOI: https://doi.org/10.12688/f1000research.6555.2

Devaraju, A, Klump, J, Tey, V, Fraser, R, Cox, S and **Wyborn, L.** 2017. A Digital Repository for Physical Samples: Concepts, Solutions and Management. In: Kamps, J, Tsakonas, G, Manolopoulos, Y, Iliadis, L and Karydis, I (eds.), *Research and Advanced Technology for Digital Libraries*. Cham: Springer International Publishing. pp. 74–85. DOI: https://doi.org/10.1007/978-3-319-67008-9_7

Digital Science, Fane, B, Ayris, P, Hahnel, M, Hrynaszkiewicz, I, Baynes, G and **Farrell, E.** 2019. The State of Open Data Report 2019. *Digital Science*. DOI: https://doi.org/10.6084/M9.FIGSHARE.9980783.V2

[6] https://curator.jsc.nasa.gov/lunar/.

[7] https://scoss.org/.

European Commission. 2017. H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 (No. Version 3.2). Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

Evangelou, E, Trikalinos, TA and **Ioannidis, JPA.** 2005. Unavailability of online supplementary scientific information from articles published in major journals. *The FASEB Journal*, 19(14): 1943–1944. DOI: https://doi.org/10.1096/fj.05-4784lsf

Fenner, M and **Aryani, A.** 2019. Introducing the PID Graph. Available at https://doi.org/10.5438/jwvf-8a66 [Last accessed 14 June 2020].

Hills, DJ. 2015. Let's make it easy: A workflow for physical sample metadata rescue. *GeoResJ*, 6: 1–8. DOI: https://doi.org/10.1016/j.grj.2015.02.007

Hodson, S, Mons, B, Uhlir, P and **Zhang, L.** 2019. The Beijing Declaration on Research Data. *CODATA*. DOI: https://doi.org/10.5281/zenodo.3552330

Hsu, C-N, Bandrowski, AE, Gillespie, TH, Udell, J, Lin, K-W, Ozyurt, IB, Grethe, JS and **Martone, ME.** 2020. Comparing the Use of Research Resource Identifiers and Natural Language Processing for Citation of Databases, Software, and Other Digital Artifacts. *Computing in Science & Engineering*, 22(2): 22–32. DOI: https://doi.org/10.1109/MCSE.2019.2952838

Kwon, D. 2019. The Push to Replace Journal Supplements with Repositories. *The Scientist*. Available at https://web.archive.org/web/20190829200838/https://www.the-scientist.com/news-opinion/the-push-to-replace-journal-supplements-with-repositories--66296 [Last accessed 21 June 2020].

Lehnert, K, Klump, J, Wyborn, L and **Ramdeen, S.** 2019. Persistent, Global, Unique: The three key requirements for a trusted identifier system for physical samples. *Biodiversity Information Science and Standards*, 3. DOI: https://doi.org/10.3897/biss.3.37334

Poirier, L and **Costelloe-Kuehn, B.** 2019. Data Sharing at Scale: A Heuristic for Affirming Data Cultures. *Data Science Journal*, 18. DOI: https://doi.org/10.5334/dsj-2019-048

Santos, C, Blake, J and **States, DJ.** 2005. Supplementary data need to be kept in public repositories. *Nature*, 438(7069): 738–738. DOI: https://doi.org/10.1038/438738a

Stall, S, Yarmey, L, Cutcher-Gershenfeld, J, Hanson, B, Lehnert, K, Nosek, B, Parsons, M, Robinson, E and **Wyborn, L.** 2019. Make scientific data FAIR. *Nature*, 570(7759): 27–29. DOI: https://doi.org/10.1038/d41586-019-01720-7

Stocker, M, Darroch, L, Krahl, R, Habermann, T, Devaraju, A, Schwardmann, U, D'Onofrio, C and **Häggström, I.** 2020. Persistent Identification of Instruments. *Data Science Journal*, 19. DOI: https://doi.org/10.5334/dsj-2020-018

Vines, TH, Albert, AYK, Andrew, RL, Débarre, F, Bock, DG, Franklin, MT, Gilbert, KJ, Moore, J-S, Renaut, S and **Rennison, DJ.** 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, 24(1): 94–97. DOI: https://doi.org/10.1016/j.cub.2013.11.014

Wilkinson, MD, Dumontier, M, Aalbersberg, IjJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and **Mons, B.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Wyborn, L, Ramdeen, S, Lehnert, K and **Klump, J.** 2020. Targeting the Bullseye of Metadata for Material Samples: Can We Define a Minimum Kernel for Transdisciplinary Interoperability? *Research Data Alliance 16th Plenary Poster*.

How to cite this article: Plomp, E. 2020. Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments. *Data Science Journal,* 19: 46, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2020-046

Submitted: 24 October 2020 **Accepted:** 16 November 2020 **Published:** 04 December 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS