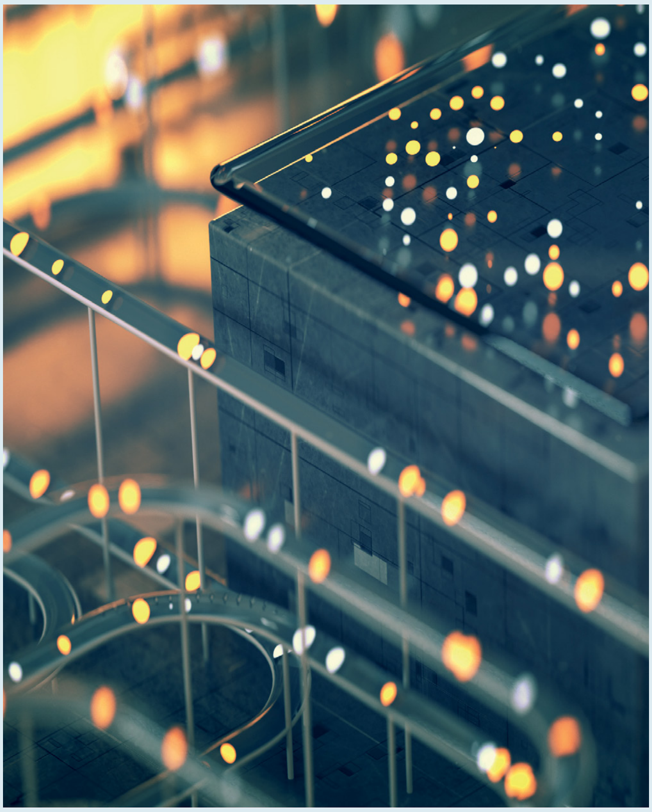
Knowledge Exchange

Case study

Persistent identifiers for research instruments and facilities: an emerging PID domain in need of coordination

February 2023



This case study is part of a series that has been produced within the study on “Risks and Trust in pursuit of a well-functioning PID infrastructure for research” commissioned by the Knowledge Exchange in July 2021. The main outcome of this study is a report examining the current PID landscape with an emphasis on its risks and trust-related issues.

This complementary series of case studies aims to provide a deeper insight into specific areas of activity, workflows and stakeholders within this wider PID landscape.

Title: Persistent identifiers for research instruments and facilities: an emerging PID domain in need of coordination

Authored by: Pablo de Castro, Ulrich Herb, Laura Rothfritz and Joachim Schöpfel (scidecode science consulting).

Email: **office@knowledge-exchange.info**

DOI: **10.5281/zenodo.7330372**

All content published is shared under a Creative Commons Attribution licence (CC BY 4.0) **creativecommons.org/licenses/by/4.0**

Contents

1. Rationale

This case study aims to explore the challenges faced and the opportunities offered by the gradual implementation of emerging PIDs.

The main focus of the case study is persistent identifiers for research instruments and facilities (PIDINSTs from now on), but the analysis is actually aimed to cover any emerging PID infrastructure and thus has links to other PID areas like persistent identifiers for conferences (ConfIDs) and – to a certain extent – to PIDs addressed in other case studies such as IGSNs for samples and ROR IDs for organisational identifiers.

With the Research Organization Registry (ROR) already up and running, the latter area of OrgIDs is clearly more advanced at the time of writing than this just-emerging PIDINST layer, but there are common challenges for both domains to come to fruition. This case study builds on the one devoted to OrgIDs.

One key aspect in this regard is the fact that the use cases for OrgIDs were identified long ago in the report "Review of selected organisational IDs and development of use cases for the Jisc CASRAI-UK Organisational Identifiers Working Group" released in 2015[1]. Since this report was released, the aim has been to identify the most suitable approach to creating an infrastructure and a governance structure that would be able to support at least some of such use cases. On the other hand, the use case analysis for PIDINSTs has not yet been made, and this is something that this case study will try to address.

Same as in the case of OrgIDs and other emerging PIDs, the largest risk perceived at the moment is that of fragmentation and subsequent lack of uptake.

This case study will hence focus on identifying the various existing initiatives exploring mechanisms to implement PIDINSTs and the multiple stakeholders that are simultaneously looking into this area with little coordination across them. The emphasis will then be on the need to figure out coordination and common awareness-raising mechanisms for the community to be able to advance together in this area.

2. The case for PIDINSTs

As mentioned in the introduction above, there are plenty of concurrent efforts going on to explore the best way to produce a PIDINST layer, but – as opposite to how the Jisc-CASRAI UK WG on OrgIDs addressed the perceived needs in this specific domain – no clear cross-stakeholder framework has been provided to date on the use cases that PIDINSTs would need to address.

Multiple stakeholders are working on the technical and community-management requirements around PIDINSTs – researchers and funders being the most prominent ones – but the case for PIDINSTs hasn't been clearly outlined yet. This is perceived as a significant risk to the uptake of this emerging PID, since a lack of awareness on the objectives the initiative aims to achieve would obviously impact the willingness of other stakeholders – institutions and publishers – to become involved in it.

A basic analysis on the relevant use cases for PIDINSTs is attempted below, but these suggestions should ideally be validated by a working group that could also serve as a much-needed coordination forum across stakeholders. First it is worth pointing out that PIDINSTs represent a much more ambitious and complex endeavour than author and organisational IDs. Some reasons for this complexity are listed below:

- It may well be possible to implement author ID, OrgID and research output ID layers without a direct intervention of researchers, but this is just not feasible for PIDINSTs. Even if these assets may be paid and operated by research funders and institutions, researchers are the truly knowledgeable stakeholder with regard to research instruments and facilities. It's no coincidence that the PIDINST WG[2] running within the Research Data Alliance (RDA) is purely researcher-driven with little if any involvement from research funders and institutions[i].
- Research instruments and facilities tend to be jointly addressed in the discussion on PIDINSTs, but these are two clearly different areas. While it may be feasible to figure out a unified approach to their persistent identification, early analyses show that facilities are much easier to address than research instruments.
- Perhaps the main reason why research instruments are much harder to address than facilities is summarised in these two overlapping questions: *"What is an instrument?"* and *"What instruments are relevant enough to be assigned a PID and which ones are not?"*
- One of the most evident differences between facilities and instruments is the variability of the landscape a PIDINST initiative would be trying to uniquely and persistently identify. While research facilities are fairly stable pieces of research infrastructure, new instruments are being purchased

i. The initial composition of the RDA PIDINST WG is available on pp. 12-13 in reference [3] below. It includes numerous representatives for data centres, supercomputing centres and other large-scale research facilities but shows little presence of research funders, research-performing organisations or the wider scholarly communications community.

and set in operation every year that would need to be identified anew. Many others are decommissioned, and this should also be reflected in a PIDINST record. The level of curation and maintenance that PIDINST records for research instruments would require is hence much more significant than in the case of facilities.

- Same as in the case of OrgIDs, it is not clearly defined who should own the PIDINST records for curation and management purposes. Institutions may be the best-placed stakeholders for instruments, while funders could naturally be the default choice for taking care of facilities, but the workflows for record curation are simply not in place at this stage.

- The case for the usefulness of PIDINSTs to the wider scholarly communications community is at the moment much weaker than any of the other PIDs under discussion. There is usually a clear-cut main goal for each PID layer – this is unique identification and subsequent disambiguation of researchers for author IDs and unique identification of affiliation for OrgIDs – but in the case of PIDINSTs this is far less clear right now.

- Furthermore, the research information workflows where identifiers for research instruments and facilities would clearly be a very useful asset to have are remarkably far from the usual scholarly communications stakeholders, meaning research libraries. This will be elaborated on in the section on use cases below, but at the moment the activity for collecting information on institutional research facilities and instruments regularly falls within the remit of research offices rather than with libraries, while discussions on metadata sets tend to involve the latter. This poses an additional internal coordination challenge that will need to be addressed.

The RDA PIDINST WG published a “case statement” in Dec 2017 making the case for a “community-driven solution for globally unique and unambiguous identification of instruments instances that are operational in the sciences”[3]. A series of use cases for these PIDs for research instruments and facilities was sketched that remains the closest attempt available to date to answer the question “Why do we need PIDINSTs?”

These use cases, which are not associated to specific stakeholders, include:

- Metrics that quantify the use of instruments and the rationale for future funding

- Link data to the instruments that generated them (provenance), improving the interpretation and validity of data

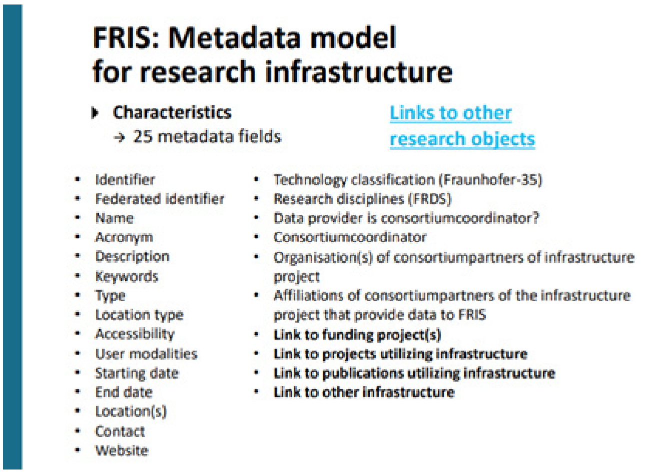
- Aid equipment logistics and mission planning

- Facilitate interoperability and open data sharing, especially in advancing technologies that foster sharing of instruments

- Improve the discoverability and visibility of instruments and their data, published on the web

The two strongest use cases on this list are the two first ones, which mainly serve the needs of research funders and institutions. Both the wish to collect usage data (including for collaborations between Academia and Industry) for often very expensive research facilities and instruments and the Open Science-aligned attempts to link research datasets to their instruments of provenance could be considered to fall within the remit of funders and institutions.

Some funders in Knowledge Exchange-member countries are in fact already maintaining national-level databases of instruments available across institutions in the country – see for instance the database equipment data promoted by the Engineering and Physical Sciences Research Council (EPSRC) and Jisc in the UK with 17,538 items (instruments) from 51 UK organisations (universities) at the time of writing[4]. The main use cases for such initiatives are to improve visibility and discoverability of such assets (item 5 in the bullet point list above) and above all to foster the sharing of such equipment hosted at specific institutions with external stakeholders. A series of posts highlighting specific pieces of research infrastructure at various UK universities was published in Jisc's "Sharing Research Equipment" blog[5] alongside the equipment.data database.



Slide from presentation "Research infrastructures: metadata model & data capturing in FRIS"[6]

More recently, the Department of Economy, Science and Innovation (EWI) of the Flemish Government has started using their Flanders Research Information Space (FRIS) system to explore the feasibility of collecting information on research equipment and facilities held at Flemish universities and beyond by proposing a common metadata set approach for their description, see a slide from a recent presentation above.

It's worth noting that although the first metadata element in this set is an identifier, there is no mention to any international, unique, persistent ID (which of course does not exist right now). These two cases do in fact highlight the risk of landscape fragmentation via national-level initiatives. On the other hand, as seen in the case study for the Dutch DAI, national-level initiatives may be a good way to start progressing while a well-established international effort is able to consolidate.

Also worth bearing in mind is the fact that these use cases for PIDINSTs contain no specific "publisher use case". This means a strong difference with the cases for author IDs and OrgIDs explored in other case studies in this series (and of course for the original Crossref DOIs for publications). In both such cases, the publisher use case was the clearest driver for the emergence of an international layer of PIDs – be it for author disambiguation in publications, or for the correct identification of their affiliations. It could be argued that it is publishers, grouped under the Crossref banner, who have pushed the hardest for the implementation of a PID layer thus far.

Publishers are also mentioned in the use case analysis this PIDINST WG has conducted, but just to state that the references to research instruments and facilities included in submitted manuscripts could also be processed by publishers based on the widespread availability of a PIDINST layer. A further element could in fact be included in manuscript submission systems to allow these references to be provided by researchers, but this is not a particularly solid driver to promote publisher involvement in such an initiative unless there was a clear requirement for this from a critical mass of research funders. DataCite is already involved in this effort around the implementation of PIDINSTs[7] and could perhaps provide some guarantees as an alternative community-driven stakeholder, especially in view of the relevance of the case for asserting dataset provenance within DataCite DOIs for datasets. It remains to be seen though whether an effort to produce an international PID layer can succeed without a direct involvement from publishers.

While publishers are not too much in the picture, a specific, somewhat unique stakeholder is repeatedly mentioned in the RDA PIDINST WG documentation – the institutional instrument database provider. Instrument manufacturers, also hinted to be a potentially important player for equipment identification purposes,

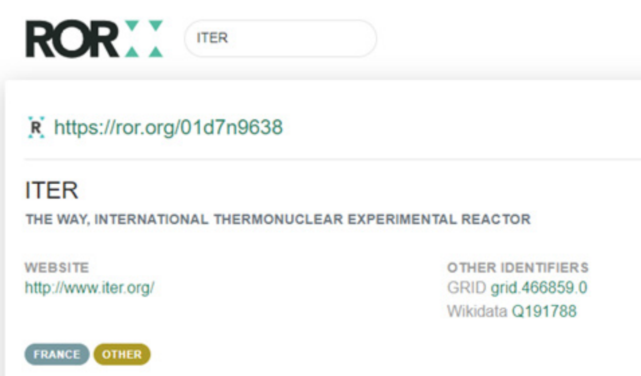
are another case-specific stakeholder the PIDINST WG work suggests to work with. These stakeholders have no equivalent in the landscape analysis for other emerging PIDs. This illustrates the significantly higher levels of complexity in the area of PIDINSTs. The institutional instrument database providers in particular are public or private initiatives working directly with universities to allow them to provide a snapshot of their available research instruments and facilities. They do not just serve the discoverability and enhanced visibility case study for institutions, but they frequently also include a booking system for requesting research time at the facilities or instruments. The Kit-Catalogue project led by Loughborough University in the UK[8] and later adopted by other HEIs in the country is a good example of such an initiative, but there are many others[ii, 9].

Another case-specific aspect worth pointing out in the PIDINST landscape is the key role played by institutional technicians in the operation and description of research instruments and facilities. These technicians are often shown as responsible for a specific piece of institutional equipment in the available metadata for these. In fact, it's these professional profiles – often overlooked for the purpose of assessing the institutional research output but increasingly brought into the picture by recent studies and reports[10] – who provide the descriptions of institutional research facilities that are being published in discipline-specific outlets like the *Journal of large-scale research facilities* (JLSRF)[11]. The journal homepage states that:

The Journal of large-scale research facilities (JLSRF) publishes articles describing large-scale scientific equipment. This covers large-scale equipment from all scientific disciplines intended to be used by scientists who are not affiliated to the institution operating the facilities (dedicated user operation). The articles in JLSRF provide scientists with a simple means to reference large-scale facilities in their publications.

ii. Interestingly, the collection of institutional instrument database provider systems in [9] makes emphasis on the case for booking instrument/facility time at the appropriate equipment as a mechanism to promote equipment sharing and not so much on the (perhaps implicit) objective of increasing discoverability and visibility for the institutional equipment.

The case statement issued by the RDA PIDINST WG cited in the bibliography suggests in fact that the DOIs issued by this journal for its published articles could be used as a sort of persistent identifier for referencing the facilities described. This approach is not fully aligned with the way persistent identifiers typically point at a metadata set describing the object (see an example for an OrgID below) and reveal the fragmented early-stage, strongly researcher-led, bottom-up attempts to identify research facilities.



Some issues identified in the analysis of the early-stage work carried out so far in the area of PIDINSTs are listed below. Some of these will be analysed in more detail in the section devoted to risks and trust at the end of this case study.

- Some types of research facilities and equipment and some countries are overrepresented in the RDA PIDINST WG, while other relevant stakeholders like research funders and universities are largely absent from the discussions. This leads to a somewhat skewed analysis of the use cases and the roadmap for PIDINST implementation.
- Partially as a result of the absence of universities in the discussions, low-hanging-fruit cases for the purpose of persistent identification of (institutional) research equipment and facilities are barely addressed in the WG reports
- Research articles summarising the results of the work carried out by this working group have only recently been published[12,13] and haven't yet had the time to influence the developments in the area even if the second reference already includes a suggested metadata schema for describing instruments and facilities.

2.1 Technical approaches to PIDINST implementation

The current PIDINST landscape shows a range of approaches towards persistent identification of research instruments and facilities, including DataCite DOIs, OrgIDs and a myriad of internal identifiers within instrument registries and databases mostly kept and maintained by research-performing organisations but also increasingly by research funders in different countries. This fragmented landscape is characteristic of an emerging PID area and doesn't represent a significant issue for the eventual implementation of an international PIDINST layer once the appropriate procedures and workflows are defined, but it stresses the need to find suitable coordination mechanisms across the various stakeholders involved in the endeavour.

The fragmented PIDINST landscape described so far resembles somewhat the attempt to implement a specific, national level author ID in the Netherlands before the arrival of international initiatives in the area like ISNI and ORCID. The case study devoted to the Dutch DAI within this series shows that the replacement and superseding of this early-stage identifier by a mix of ISNIs and ORCIDs posed no major issues in the evolution of the PID landscape in the country.

This section will analyse and provide examples for the various identification mechanisms that are currently being used to generate PIDINSTs. Some suggestions will also be made to more effectively achieve a more widespread – perhaps in terms of reaching a critical mass – PIDINST implementation.

DOIs are the most straightforward choice for persistent identification of research instruments and facilities. They not only provide a well-established, well-functioning infrastructure and governance layer, but DOIs minted by DataCite are also fully aligned with the main case for PIDINSTs, namely tracing the provenance of research datasets also identified by means of DataCite DOIs. But as we have seen for the ITER-Cadarache example above, the choice for DOIs is far from being the only approach, especially for large-scale research facilities operating as standalone research equipment outside institutions.

As a non-profit, community-driven member organisation, DataCite is playing the PID infrastructure provider role for these early-stage PIDINST implementations[7]. This is partly a result of the collaboration networks established in the course of the EU-funded FREYA-777523 project ("Connected Open Identifiers for Discovery, Access and Use of Research Resources", 2017-2020). The section in the FREYA project website devoted to "prototypes of new PID services"[14] includes two subsections on PIDs for scientific instruments and PIDs for research facilities. Project partners PANGAEA and the UKRI Science and Technology Facilities Council (STFC) are referenced as frontrunners in the implementation of PIDs in each of both areas.

The DOI-based persistent ID layer for a research instrument mentioned as an example in the DataCite blog post at [7] is for one specific seismic network – the Shumagin-East Aleutian Network in the Alaska Peninsula, https://doi.org/10.7914/SN/SH – within the International Federation of Digital Seismograph Networks (FDSN), see figure below. The landing page for this DOI shows a loosely structured metadata collection including among others a description of the network, its geolocation, a suggested citation to be added to manuscripts, the academic institution in charge of its operation and a list of the stations that make up the network. Two elements highlighted in the figure below are the DOI itself and a link that allows the operator of the network to update the information on the page where appropriate.

While this is quite a complex instance of a research instrument, the feature enabled by the FDSN to allow the operating institution to curate and maintain the record for the facility is a valuable example for a curation workflow that could be applicable to other emerging PIDs too.

FDSN International Federation of Digital Seismograph Networks

Home / Networks / SH: Shumagin-East Aleutian Network   Sign in

SH: Shumagin-East Aleutian Network

FDSN Network Information

Are you the operator of this network? Update this information.

FDSN code	SH	Network name	Shumagin-East Aleutian Network (Shumagin Net)
Start year	1981	Operated by	Lamont Doherty Earth Observatory (LDEO), Columbia University ROR
End year	1991	Deployment region	
Description	Analog-telemetered network in the Alaska Peninsula, Shumagin Islands and East Aleutians. First stations were installed late 1960's with major deployment mid-1970's. Digital recording data is available from 1982 - 1991, when network was decommissioned. This is a rare network directly above an active subduction zone megathrust		

Citation Information

Digital Object Identifier (DOI)	https://doi.org/10.7914/SN/SH
Citation	Lamont Doherty Earth Observatory (LDEO), Columbia University. (1981). *Shumagin-East Aleutian Network* [Data set]. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/SH For more: DataCite (JSON | XML | BibTeX)
Citation notes	Two refereed citations: Abers, G. A. (1992), Relationship between shallow- and intermediate-depth seismicity in the eastern Aleutian subduction zone, Geophys.Res. Lett., 19, 2019-2022. Reyners, M., and K. Coles (1982), Fine structure of the dipping seismic zone and subduction mechanics in the Shumagin Islands, Alaska, J. Geophys. Res., 87, 356-366.

Although the process is not complete and plenty of networks do not yet have a DOI, the FDSN has created further DOIs for other seismic networks within the federation such as the ones listed below. Note that the authors' names on the citations are for the operators of the specific network, following the practice established by the JLSRF at [11].

- Erin Pettit, Ted Scambos, & Martin Truffer (2015). ***RAPID: Observing the Disintegration of the Scar Inlet Ice Shelf***. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/1T_2015
- Heit, B., Yuan, X., Almendros, J., Abella, R., Carmona, E., Aguí, F., & Carrión, P. (2020). ***BRAVOSEIS Onshore Seismic Array.*** GFZ Data Services. https://doi.org/10.14470/0Z7563857972
- Alex Brisbourne, & Andy Smith (2016). ***Bed Access, Monitoring and Ice Sheet History (BEAMISH, 2016-2019)*** [Data set]. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/9B_2016

As mentioned above, PANGAEA was a FREYA project partner in the domain of PIDs for research instruments and facilities and is thus a frontrunner in the implementation of PIDINSTs, see for instance the DOI for the Polarstern research vessel stemming from an article devoted to this research facility published in the JLSRF:

Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung (2017). ***Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute.*** Journal of large-scale research facilities, 3, A119. http://dx.doi.org/10.17815/jlsrf-3-163

Moreover, datasets hosted by PANGAEA are already using the reference to both the research facility and the research instrument from which the data was obtained, see

Witte, Hannelore (2018). ***Processed 2 minutes-averaged continuous VM-ADCP (vessel-mounted Acoustic Doppler Current Profiler) profiles during Polarstern cruise PS82.*** Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, https://doi.org/10.1594/PANGAEA.887545

The metadata set for this persistently identified PANGAEA dataset includes links to the Polarstern research vessel described at the JLSRF, to the specific campaign (PS82) via a TIB-hosted DOI-based ID and also to the specific research instrument that was used on the vessel, Acoustic Doppler Current Profiling (ADCP), TRDI Ocean Surveyor, 153.6 kHz, https://doi.pangaea.de/10013/epic.47834.d001. This link gives access to a 2-page instrument description by the manufacturer Teledyne RD Instruments in California.

It is no coincidence that – same as for the IGSN emerging PID addressed in another case study in this series – both examples for early-stage implementation of PIDINSTs belong to the field of Earth Sciences. It is this area, with its large data banks[15], where the case for linking datasets to their provenance is most acutely felt at the time. At the same time, these are fairly sophisticated examples of persistent identification for instruments and facilities, and it's worth wondering how these early-stage practices may be taken over by more down-to-earth instances like universities. A suggestion in this regard is made in the following section.

Another very relevant research discipline for the implementation of PIDINSTs is that of high-energy physics, and specifically the trans-European community of national Neutron and Photon facilities. The well-established collaborative network across these facilities and institutions formally started with the PaNdata ODI-283556 FP7 project in 2011 ("Photon and Neutron Data - Open Data Infrastructure", 2011-2014). The main project goal was to establish a data infrastructure, federated among the European Neutron and Photon facilities, to enable the scientific communities access, analysis and sharing of scientific data in a collaborative research environment.

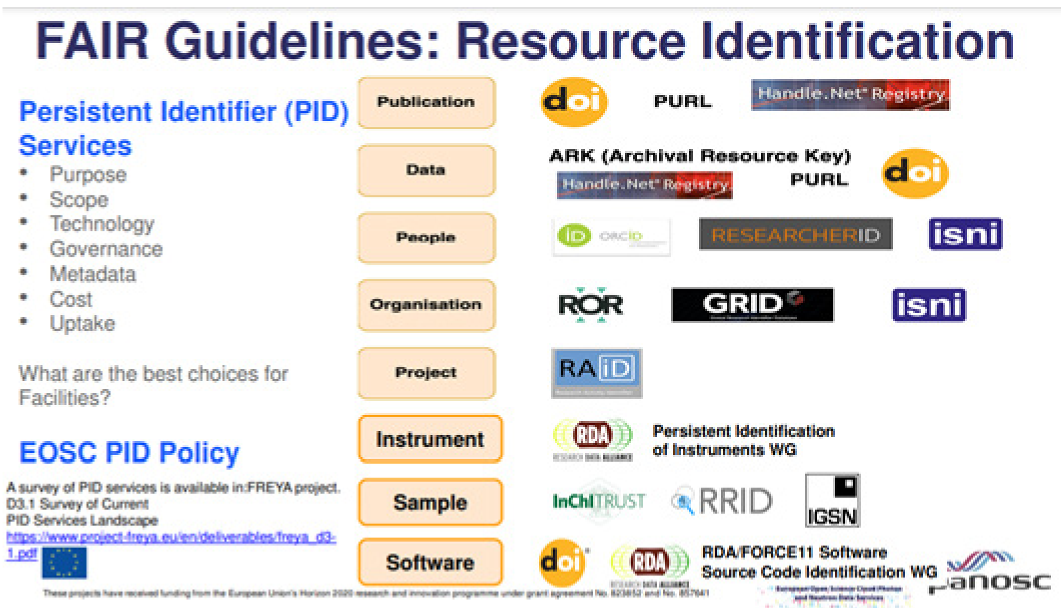
Building on this collaborative network, this consortium of facilities in the high-energy physics discipline are now running two H2020-funded projects under the umbrella of the European Open Science Cloud (EOSC), namely PaNOSC-823852 ("Photon and Neutron Open Science Cloud", 2018-2022) and ExPaNDS-857641 ("EOSC Photon and Neutron Data Services", 2019-2023).

As shown on the figure below, taken from a Nov 2020 PaNOSC/ExPaNDS project presentation[16], persistent identification of research facilities and instruments is very much on the radar of this research community under the concept of "FAIR research infrastructures". The PaNOSC/ExPaNDS network has in fact recently held specific workshops on the topic of PIDs for research facilities[17]. Same as in the case for the early-stage PIDINST implementation in the Earth Sciences above, this 'niche' HEP research network is detached from the mainstream scholarly communications community at universities. The way these efforts are being conducted suggests however that the EOSC and specifically its PID Policy and Implementation Task Force could play a significant role in the promotion of PIDINSTs via the various EOSC science demonstrators in multiple disciplines that were carried out under the EOSCpilot project.

The examples for PIDINSTs explored so far are all based on DOIs as a persistent identifier type. However, an increasing number of research facilities are also getting top-level OrgIDs in a parallel process. The ROR ID for ITER-Cadarache has been shown earlier in this case study, and other facilities with associated ROR IDs include:

- the Diamond Light Source at https://ror.org/05etxs293 or the SKA Observatory, https://ror.org/01pwgd096, in the United Kingdom,
- the German Electron Synchrotron (DESY) in Germany, https://ror.org/01js2sh04,
- the European Spallation Source in Sweden, https://ror.org/01wv9cn34,
- the Barcelona Supercomputing Center (BSC) at https://ror.org/05sd8tv96 or the Large Canary Islands Telescope, https://ror.org/03efc2j36 in Spain.

This might suggest diverse options for a persistent identification of research instruments and facilities, especially bearing in mind the approach described in the case study devoted to OrgIDs whereby multiple-level ROR IDs could potentially be used to assign PIDs to instruments held within these large-scale research facilities.



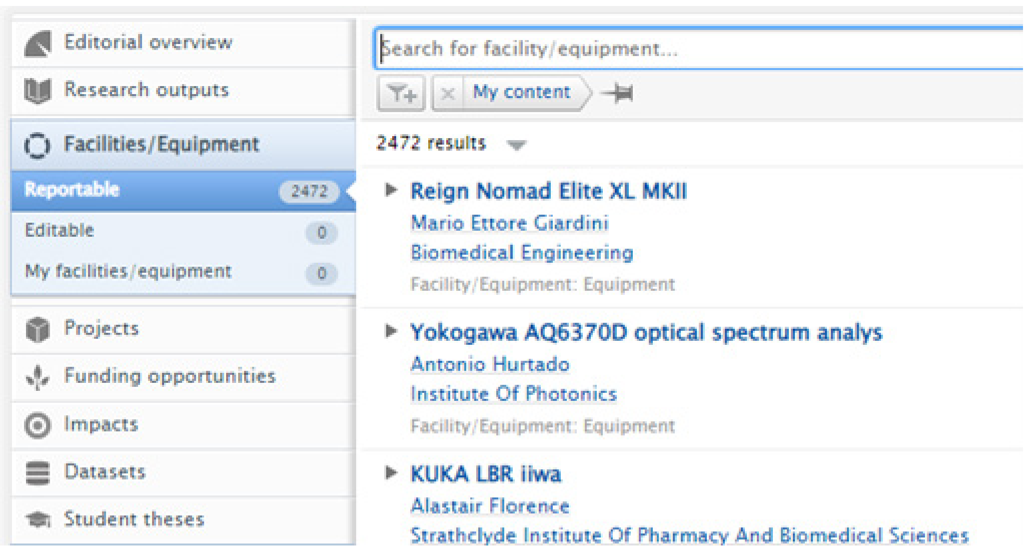
2.2 The role of Current Research Information Systems (CRIS) for the implementation of PIDs

As shown in the Directory of Research Information Systems (DRIS) maintained by the non-profit member association euroCRIS[18], there are over a thousand operational Current Research Information Systems worldwide devoted to capturing all aspects of the research activity carried out at research-performing organisations, mostly universities. These systems are widespread in most European countries and particularly in the six Knowledge Exchange member countries.

CRIS systems, also known as Research Information Management Systems (RIMS), often operate on the basis of the CERIF standard (for Common European Research Information Format) for metadata description purposes[19]. CERIF is composed of a network of interconnected entities, with persons (researchers), organisations and projects at its core. Research equipment is a secondary-level entity in the CERIF data model, and it is linked to the core ones and to various other entities. This is essentially the same concept as the research graphs for describing research and specifically the same as PID graphs using persistent identifiers to establish such interconnections across entities. The recently released [DataCite] Metadata Schema for the Persistent Identification of Instruments[13] is in fact well-aligned with the structure of the CERIF cfEquipment entity.

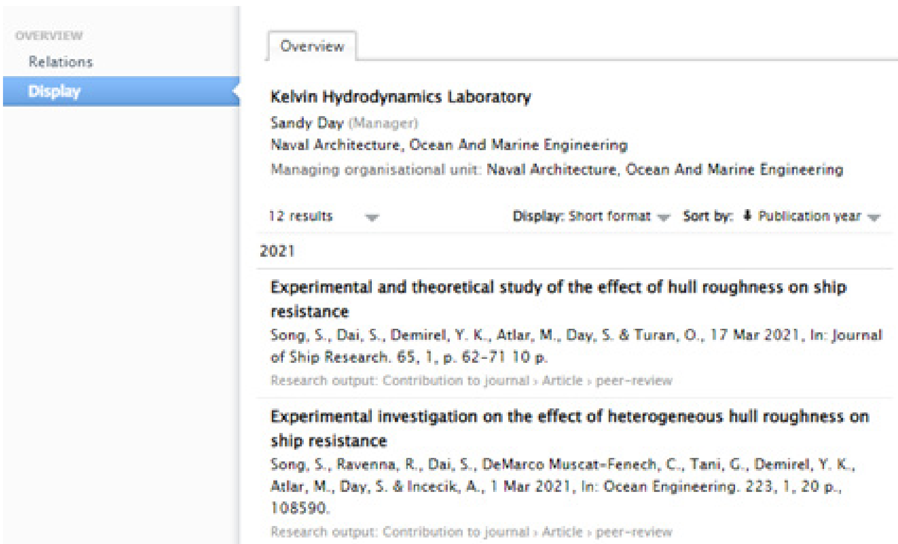
CRIS conferences and membership meetings held by euroCRIS in the past have seen contributions in the area of research information management via persistent identifiers and specifically on collection of information on research instruments and facilities, see for instance [20] and [21]. Because these events are usually attended by CRIS managers both for institutional and national/ regional CRIS, they could represent a good opportunity to make the case for the implementation of emerging PIDs such as PIDINSTs, to discuss the associated workflows and to exchange good practices in the area.

Moreover, institutional CRIS systems very often contain records for institutional research equipment, see an example in the figure below for the list of instruments and facilities available at the institutional CRIS at the University of Strathclyde in Glasgow, UK. The CERIF data model underpinning these institutional CRISs already allows publications and datasets to be linked to instruments and facilities available in the institutional database – even if the provenance equipment does not yet have persistent identifiers associated with it. The UUID structure underlying these CERIF entities already allows these links to be set.



As a result of this, PIDINSTs for the available equipment could quickly be made available by using the DataCite services for DOI minting that member institutions are already using for minting DOIs for datasets. However, this would be a largely pointless exercise if the new PIDINST layer were not underpinned by a default use case for using such persistent identifiers to accurately reference instruments and facilities in – for instance – manuscripts submitted for publication and their underpinning datasets.

This need for services to be developed on top of an eventual PIDINSTs layer suggests research funders might be the key driver for their adoption. By requiring this additional feature – same as ORCIDs were required in past iterations of the gradual emergence of a PID landscape – funders would be able to promote the citation of instruments and facilities by researchers in their research publications and datasets. Funders would also be able to obtain usage metrics for the research instruments and facilities they are funding at institutions. Publishers could also explore the required mechanisms to include references to PIDINSTs in manuscripts and journal articles. This means a significant cultural change, one that would require researchers to apply a systematic approach to citing PIDINSTs, but this could potentially be supported by the gradual inclusion of the maintenance and operation of institutional research instruments and facilities within evaluation frameworks for universities like the Knowledge Exchange Framework (KEF) in the UK.



3. Issues around risks and trust regarding the implementation of PIDINSTs

Serious risks associated with the implementation of PIDs for research instruments and facilities are perceived to be the fragmentation and lack of coordination across initiatives working in the area, which could in turn result in an uneven uptake. This is a common trait to all emerging PIDs, but in the case of PIDINSTs, the landscape fragmentation does not only involve significant methodological variations across research disciplines, but also the fact that whole stakeholders within the scholarly communications community are currently unaware of the recent progress in this domain. In order for this sort of PIDs to consolidate – meaning not just PIDINSTs but also the additional 'technical' PIDs – putting them firmly on the radar of actors like research funders and institutions is seen as a must.

The very early-stage implementation of PIDINSTs at the time of writing raises a good number of potential risks and trust-related issues. Some of these – such as the risk of fragmentation and lack of uptake by the wider community – are common to any emerging PID. More specifically, PIDINSTs are the first example in this series of case studies for what we call a 'technical PID', meaning that its implementation has so far mainly being driven by researchers and research facility managers with little involvement from other stakeholders like institutions or research libraries (ISGNs for geo samples are another example for 'technical PIDs' addressed in this series). This is in contrast to the more 'admin-oriented PIDs' whose use cases much more clearly serve the objectives of the wider scholarly communications community. These are not clear-cut categories, but given that the risks are common for PIDs included in any of such groups, it's worth looking into this classification in a bit more detail.

3.1 'Technical' vs admin-oriented PIDs

There is a dichotomy within the current and emerging PID landscape between what we could call "technical" and "admin-oriented" identifiers. Technical PIDs are promoted as bottom-up workflows by researchers who perceive the need for the persistent identification of objects they're regularly working with, be it geo samples (IGSN), research equipment and facilities (PIDINSTs) or clinical trials (ISRCTNs). Admin-oriented PIDs on the other hand (such as – among others – ORCIDs, OrgIDs and the emerging ConfIDs) are implemented in a more (if not completely) top-down fashion by a range of stakeholders that do not include researchers – typically institutions, publishers and research funders – in order to introduce some much-needed standardisation in the scholarly communications landscape for research information management purposes.

The workflows for the implementation of these admin-oriented PIDs tend to be much more coordinated across countries than the usually rather fragmented attempts to promote technical PIDs. It is this international coordination what allows researchers to eventually use and benefit from the admin-oriented PIDs too, but at the moment there's no clear roadmap for the adoption of the technical PIDs by the stakeholders in the scholarly communications domain. The community management role that actors like DataCite and Crossref may play in this regard is seen as critical. However, there's still some way to go in terms of making the case for the usefulness of technical PIDs so that stakeholders sitting on the admin-oriented PID side of the landscape may be persuaded to promote and support these other technical PIDs as well. Without such coordination between both sides of the landscape, the risk of fragmentation and lack of uptake will remain paramount and may well prevent the emergence of a widespread PID layer for technical PIDs due to the sheer lack of awareness.

3.2 Community engagement

The very successful international implementation of ORCID as a default PID for researchers has made it a certain blueprint for other PID implementation initiatives to follow. This would mean an integrated approach to persistent identification of a specific range of objects by the many different actors that make up the scholarly communications community. However, different PIDs could also follow different strategies for becoming widely adopted.

> Things like datasets and instruments – that's information we [funders] are not necessarily capturing as part of the digital application. Details that may be produced as part of one of the outputs and where these are being stored. That information may not necessarily come back to us, it may be deposited with another organisation and then we [would] have to connect to it. So I think there are even greater technical barriers to implementing those smaller PIDs. We are not going to be anywhere close to becoming the central depository for all PIDs, at least currently. Of course, we may never be that. There's certainly no roadmap for that. And so it may be a provider holds all the data sets, a different provider holds all the equipment and instruments, a provider holds the policy outputs, for example, if that becomes one additional area. And [as funders] we have to find a way to connect that information ourselves, but not everyone's in the position to be able to do that.

This said, it is clear at this point that there is a high risk of fragmentation and lack of uptake in the PIDINST domain. There are many parallel initiatives underway which overlap with each other with little awareness of the progress made by other stakeholders, and there's a clear need for mechanisms to bridge the gap between research communities and stakeholders. This initiative by the Knowledge Exchange to analyse the current PID landscape and explore its risks and trust-related issues could be of much help in this regard, but further efforts are clearly required to achieve a widespread adoption and subsequent use of PIDINSTs.

Part of this pending work could be conducted by organisations like DataCite, the EOSC or euroCRIS by organising events where the current status of a given PID domain could be explored with presentations from the pioneering initiatives[iii]. The purpose of these events would be to raise awareness and bridge the gap across communities, fostering a wide-ranging discussion where actors like research funders and institutions could consider the potential benefits of joining the PIDINST bandwagon. It's worth pointing out in this regard that several countries have embryonic-stage databases of research infrastructure for admin purposes that could perhaps be upgraded to

iii. An open webinar on persistent identification of instruments was held on May 19th, 2022 by the RDA PIDINST WG at https://www.rd-alliance.org/PID-instruments-May2022_webinar. This is very much along the lines of the dissemination initiatives suggested here.

PIDINST-supported assets and turned into the basis for well-accepted practices for instrument citation in datasets and publications.

A specific recommendation in this regard would be for a research funder in a Knowledge Exchange member country where early-stage PIDINST adoption initiatives are relatively consolidated (such as the United Kingdom or Germany) to consider a funding call for a project involving not just the pioneering, discipline-specific frontrunner initiatives but also other actors such as universities, institutional instrument database providers and even instrument manufacturers. The main goal of such a project would be to create new cross-stakeholder collaborative networks and to explore the workflows for the extension of early-stage PIDINST adoption practices onto more mainstream actors such as universities. It would also be for this kind of diverse project consortium to deliver a comprehensive description of the various use cases for PIDINSTs – a case currently somewhat biased due to the lack of direct input from Higher Education Institutions and research funders.

> We've also had various conversations around instruments and we do support that, at a very basic level at this stage. But yeah, also something to consider. So I would say, my advice, because I can't give you an exhaustive list, would be to assess the use cases that are needed, or what use cases you are trying to address. And then do an analysis of what is available there and then overlay that with the trust or maturity of those services and communities that would offer those. And that's what we are trying to do in partnering and working with different organisations, to elevate the trust and maturity of these services and scale more efficiently.

3.3 Technical divergence

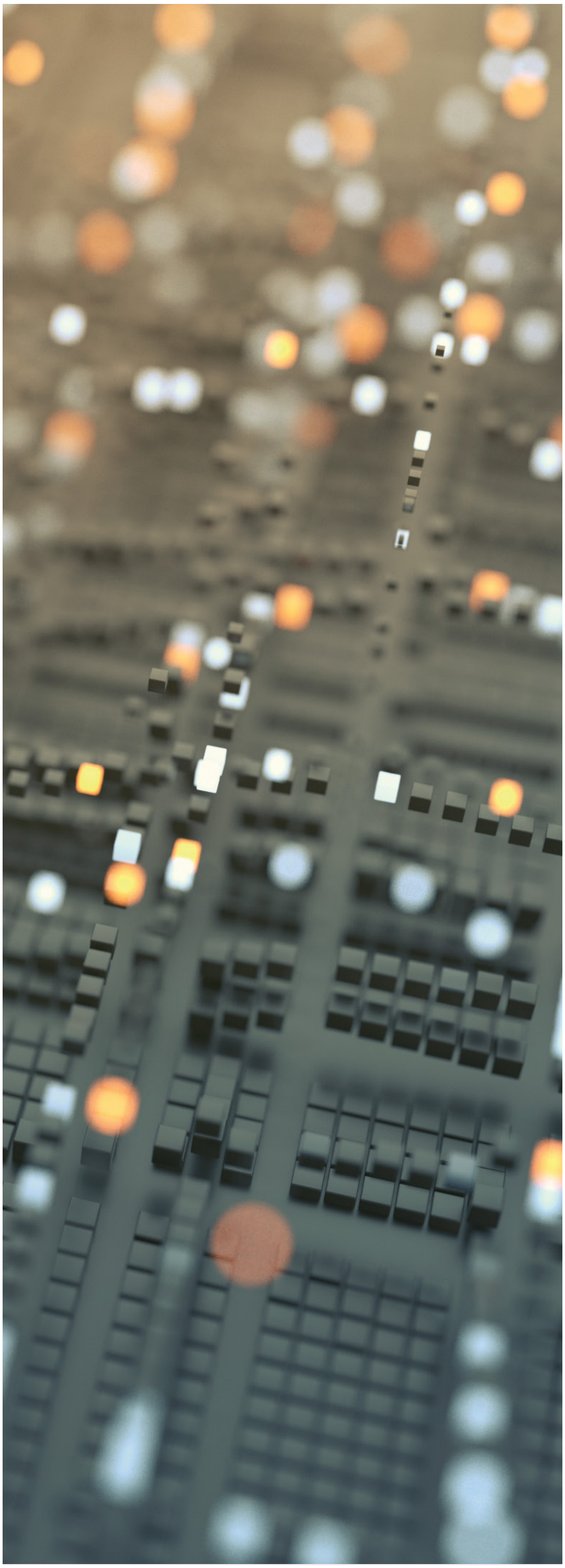
Due to the siloed nature of these early-stage, discipline-specific attempts at adopting PIDINSTs, a risk of technical divergence is presently looming over the whole domain. Several examples have been provided above on how different initiatives are using different PID standards such as DataCite DOIs, ROR IDs and DOIs for articles on research facilities published in specific journals as the basis for the persistent identification of research instruments and facilities. Part of the issue is due to the perhaps overambitious objective of addressing facilities and instruments as a single area, where these two can be seen as different domains to be independently addressed. This issue has also surfaced in the case study on OrgIDs, where competing standards such as ROR and Ringgold are due to coexist, and may not be a critical one as long as there's an overarching strategy to bridge the gaps across technical standards, but PIDINSTs are currently at a much earlier stage of implementation and this risk is seen to be more acute than in the case of OrgIDs. This is in fact a wider issue for emerging PIDs that will also affect grantIDs, an area where early-stage approaches are mainly being supported by Crossref while RAIDs are simultaneously coming into the picture too.

> The [effort around] instruments is an interesting one, because something that we [publishers] very much focus on is scalability and consistency of approaches. Now, instrumentation is very, very different from field to field and the tracking of that, I mean, we look at astronomy, they've done it incredibly well, for 20-odd year. That gives them a great consistency of data and understanding. But then if we flip over into something like some of the chemical sciences, it becomes very variable within sub-disciplines of chemistry. And so, if we could get to a place where there was some kind of identifier that worked across disciplines, it would be very beneficial, especially for abstracting, understanding of impact, identifying the ongoing types of research and trends and so on. I think the likelihood of that happening is much lower than in something like institutions or research objects or so on, because the variability is so high. Yeah, so I think it would be beneficial, but it's more unlikely.

In a way these risks of technical divergence are part of the issues around community engagement mentioned in the previous section and could probably be addressed as part of a more general awareness-raising exercise. It is important in any case to bear them in mind in order to avoid fragmentation.

3.4 Overload of PID infrastructure providers

Emerging PID domains like OrgIDs, PIDINSTs, IGSNs, grantIDs and ConfIDs are likely to be simultaneously developed and matured. The fact that the underpinning PID infrastructure and the community coordination for all these initiatives sit with a single actor in the community – namely DataCite – is also seen as a significant risk. This would be about an overload for the services – especially in the area of community engagement and management – to be provided by a single organisation. This risk could result in a slowing down of the progress in the implementation of these emerging PIDs and the inability to counter the natural fragmentation that tends to arise from early-stage efforts. As mentioned above, the participation of other stakeholders like the EOSC in the efforts to define a standardised, coordinated roadmap towards a widely implemented PID layer may well counter this risk of overload. Projects conducted under the EOSC umbrella like FAIRsFAIR and the recently started FAIR-IMPACT and FAIRCORE4EOSC ("Developing a set of EOSC-Core components to enable a FAIR EOSC ecosystem", 2022-2025) led by CSC in Finland with DataCite among its partners do in fact mean an opportunity to jointly address this challenge of pushing for the simultaneous implementation and consolidation of many different emerging PIDs.



4. Authorship

This case study has mainly been written by Pablo de Castro (University of Strathclyde and euroCRIS, ORCID https://orcid.org/0000-0001-6300-1033) within a team of consultants including Ulrich Herb (Saarland University, ORCID https://orcid.org/0000-0002-3500-3119), Laura Rothfritz (Humboldt University Berlin, ORCID https://orcid.org/0000-0001-7525-0635) and Joachim Schöpfel (University of Lille and euroCRIS, ORCID https://orcid.org/0000-0002-4000-807X) under the umbrella of scidecode science consulting (ROR https://ror.org/02c0bjd31). The work has been overseen by the Knowledge Exchange Task & Finish Group whose composition is listed at https://www.knowledge-exchange.info/event/pids-risk-and-trust.

5. Literature

1. Ferguson N, Moore R, Schmoller, S (2015) "Review of selected organisational IDs and development of use cases for the Jisc CASRAI-UK Organisational Identifiers Working Group". https://repository.jisc.ac.uk/id/eprint/5853
2. Research Data Alliance Persistent Identification of Instruments Working Group (PIDINST WG), https://www.rd-alliance.org/groups/persistent-identification-instruments-wg
3. RDA PIDINST WG (2017) Case Statement, https://www.rd-alliance.org/sites/default/files/case_statement/rda-wg-pidinst-case-statement.pdf
4. Jisc/EPSRC database equipment.data, https://equipment.data.ac.uk/status
5. Ingram C (2018) Jisc "Sharing Research Equipment" blog, https://equipment.jiscinvolve.org/wp/
6. De Bal I, Poelmans H, Dengis P (2021) "Research infrastructures: metadata model & data capturing in FRIS". Autumn 2021 euroCRIS webinars, http://hdl.handle.net/11366/1867
7. Buys M, Dasler R, Fenner M (2020) "PIDs for instruments: a way forward". DataCite blog, https://doi.org/10.5438/tdk2-2g94
8. King M (2015) "Kit-Catalogue® - the open source equipment catalogue application". http://www.efficiencyexchange.ac.uk/wp-content/uploads/KitCat_casestudy_ee_Feb15.pdf
9. Duca D (2017) "Five research equipment booking systems reviewed". Efficiency Exchange blog, http://www.efficiencyexchange.ac.uk/11219/managing-research-equipment-review-booking-systems/
10. UKRI Research England (2021) "The Role of Technicians in KE: An explorative study", https://www.mitalent.ac.uk/write/MediaUploads/PDFs/TALENT_Technicians_and_Knowledge_Exchange_web.pdf
11. Journal of large-scale research facilities (JLSRF), https://jlsrf.org/index.php/lsf
12. Stocker M et al (2020) "Persistent Identification of Instruments". Data Science Journal, 19(1): 18. DOI: http://doi.org/10.5334/dsj-2020-018
13. RDA PIDINST WG (2022) "Metadata Schema for the Persistent Identification of Instruments", https://doi.org/10.15497/RDA00070
14. FREYA project (2020) “Prototypes of new PID services”, https://www.project-freya.eu/en/pid-graph/prototypes-of-new-pids
15. British Antarctic Survey UK Polar Data Centre (UK PDC) Discovery Metadata System, https://data.bas.ac.uk/
16. Matthews B (2020) "Achieving a Photon and Neutron community federated cloud in EOSC: Enabling our facilities to produce FAIR data". ExPaNDS-EGI Workshop on FAIR Data, https://indico.egi.eu/event/5000/contributions/14511/attachments/13306/16179/ExPaNDS-EGI-workshop-FAIRData.pdf
17. ExPaNDS project (2021) "Persistent Identifiers for Research Facilities Workshop", https://www.panosc.eu/events/persistent-identifiers-for-research-facilities-workshop/
18. euroCRIS Directory of Research Information Systems (DRIS), https://dspacecris.eurocris.org/cris/explore/dris
19. Joerg B. About Common European Research Information Format. European Commission Joinup, https://joinup.ec.europa.eu/collection/eu-semantic-interoperability-catalogue/solution/common-european-research-information-format/about
20. Cox A, Gutteridge C (2016) "equipment.data – Delivering a data autodiscovery infrastructure". CRIS2016: 13th International Conference on Current Research Information Systems (St Andrews, June 9-11, 2016), http://hdl.handle.net/11366/500
21. Tatum C, Brown J (2018) "Principles and Pragmatics of “as open as possible”: persistent identifiers as the interface between research information commons and closed systems". CRIS2018: 14th International Conference on Current Research Information Systems (Umeå, June 13-16, 2018), http://hdl.handle.net/11366/658

Knowledge Exchange
C/O Jisc
4 Portwall Lane,
Bristol, BS1 6NB
United Kingdom

T +44 203 697 5804
E office@knowledge-exchange.info