RDA WG Persistent Identification of Instruments (PIDINST WG)

Case Statement

WG Charter

The Persistent Identification of Instruments RDA Working Group (PIDINST WG) seeks to propose a community-driven solution for globally unique and unambiguous identification of instruments *instances* that are operational in the sciences.

In her recent book, entitled “Big Data, Little Data, No Data” [1], Christine Borgman writes “To interpret a digital dataset, much must be known about the hardware used to generate the data, whether sensor networks or laboratory machines.” Borgman further highlights that “When questions arise [...] about calibration [...], they sometimes have to locate the departed student or postdoctoral fellow most closely involved.” This is a striking account for the role information about instruments plays in science and the costs of not being able to find and access such information.

The need to uniquely identify an instrument instance is rapidly growing in many research communities. Indeed, persistent identifiers enable unambiguous reference to digital representations of instruments, which has many potential benefits:

- 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.

Currently, there is no universal way to identify instrument instances. As the primary outcome, PIDINST WG contributes to establishing a cross-discipline, operational solution for the unique and lasting identification of active and decommissioned instruments. This case statement outlines the work planned for PIDINST WG.

Issues to be addressed

- Instruments as physical entities - What is an instrument? Implications of identifying the instrument instance as a physical object versus identifying a digital information object (metadata) about the instrument. What do instruments produce, their real-world configurations, their relations to platforms and deployments, and the implications of instrument modifications to identification (new versions).
- Granularity - Instruments can be parts of other (compound) instruments. For example, instruments can be manufactured with multiple bespoke sensor components, such as modular weather stations that simultaneously measure multiple meteorological variables. The granularity at which to reference and describe instrument instances (compound versus component) can vary for different stakeholders. How can these types of instruments be described in a generic way.
- Use cases - Support the analysis of community requirements and inform the work carried out by PIDINST WG.
- Metadata - Explore the types and sources of metadata that could be resolved under a PID and the difference between metadata registered at PID infrastructure provider (e.g. DataCite, ePIC, Crossref) vs. metadata at institutional instrument database provider. Develop a minimum common metadata schema for the registration of instruments with PID infrastructure providers.
- Machine readability, interoperability, and provenance - Investigate the need and the requirements involved to make metadata (at the institutional level) machine readable and compatible with existing interoperable technologies. Provenance, in particular the relation between data and instruments that generated them, is another aspect to be addressed.
- Landscaping - Explore the links, potential relationships and overlaps with instrument manufacturers, institutional instrument database providers, RDA groups and PID infrastructure providers.

Outcomes

The work of the PIDINST WG will contribute to the following outcomes. Note that these are long-term outcomes this WG aims at contributing to. This WG will not build a sustainable infrastructure for the persistent identification of instruments. It will merely contribute to specifying such infrastructure. The concrete deliverables of this WG are presented in the Work Plan.

- A sustainable infrastructure will support the registration of instrument instances by submitting metadata about them and allowing for minting an instrument instance PID. The PID must follow agreed standards for persistent identifiers, e.g. long-lasting actionable, descriptive digital identifiers.

- Improved understanding within research communities for how to describe instrument instances, including relations to other entities such as instrument model (type) or instrument deployment, the issue of identifying physical objects versus digital representations, and other related issues.

- Collaborations with one or more PID infrastructure provider interested in implementing the approach to persistent identification of instruments proposed by the PIDINST WG.

- Strong linkages to the activities of the RDA PID IG and other related RDA groups.

Value Proposition

Persistent identifiers provide a permanent reference to digital content that is resolvable over the internet, including a digital description of a physical object. Once assigned to an instrument instance, a PID will only ever apply to that particular instance - also beyond the instrument instance's lifetime. Therefore, communities will be able to unambiguously identify instrument instances in their own applications or management systems, making it easier to manage assets. When assigning a PID, the registrant links metadata with a level of detail appropriate to identify a distinct entity within the PID system. Using a common set of metadata will make it easier to share and collate information across disparate networks such as research e-infrastructures and make it easier to automate processes. A PID resolves to a landing page containing more detailed information about an instance, providing a permanent link to attributes such as its characteristics and capabilities, which may evolve over time.

Individuals, communities and initiatives that will benefit

The key beneficiaries are stakeholders interested in using persistent identifiers for instruments, specifically:

- *Researchers:* Persistent identification of instruments will enable linking datasets, and metadata describing them, with the instrument instances involved in measuring parameters. Unambiguous linking, additional attributes to PIDs, and PID resolution onto landing pages will provide researchers contextual information needed to determine how to appropriately process or interpret data.
- *Data repositories:* Data repositories already capture information about devices in metadata of published datasets. For instance, the metadata[1] of a dataset published by PANGAEA includes information about the device type used to measure temperature and other parameters, namely "CTD, SEA-BIRD SBE 911plus". PANGAEA links this

[1] https://doi.pangaea.de/10.1594/PANGAEA.807548

information to the corresponding vendor web page[2]. In the future, such linking could be through PIDs and at the granularity of the device instance, rather than the type. A similar case can be found at GEOFON, and other seismological data centres. GEOFON offers a web service (FDSN station-WS) that provides metadata for datasets. These metadata include information to convert data into physical units, and a description/specification of the instrument types that recorded the data, but usually not the instrument instance. The metadata follows an international standard format called StationXML, which will most probably include in its future version the possibility to add persistent identifiers to relate to entities external to the metadata content.

- *Registries*: There already exist registries that support the lookup of descriptions of entities related to instrumentation. The ESONET Yellow Pages[3] and the NERC Vocabulary Server[4] are examples of registries for sensing device types (models) or terms more generally, and their descriptions. Such registries would benefit from persistent identification of instruments since the identifiers can, e.g., be crosslinked with identifiers representing entities related to instrument instances, e.g. instrument type.
- *Hardware curators:* Facilities, such as the National Oceanography Centre's National Marine Equipment Pool (NMEP), are responsible for keeping track of an institution's instruments (or assets). For instance, the NMEP is responsible for the procurement, maintenance, calibration and provision of equipment for research expeditions as well as for providing crucial information for international custom authorities and insurance purposes during shipping. Persistent identification will help manage such assets more efficiently.
- *Manufacturers:* In addition to individual researchers, commercial manufacturers and other institutions that manufacture instruments may be agents that register instruments "at birth". In some cases, manufacturers may also maintain descriptive landing pages, and thus the mapping of instrument PID with instrument landing page. Involvement of (commercial) manufacturers can benefit these organizations by actively involving them in downstream use of instruments as well as increased visibility and use of their services.
- *PID infrastructure providers:* By adopting the outcomes of this WG, PID infrastructure providers can extend their service offering into a new object type.
- *Funding bodies and institutions:* Persistent identification of instruments is a building block towards attributing data output to instruments and assessing the contribution of instruments to science. Metrics for the "performance" of instruments ("intense of use") will support funding bodies in allocating resources and allow institutions to demonstrate the value of their instrument assets.
- *Data management service providers:* Providers of semantic sensor web publications require resolvable identification of sensor descriptions that are universally unique in order to support a global network of sensors. This was identified as a priority in a EU report [3] on data management and standards for sensors.

[2] http://www.seabird.com/sbe911plus-ctd
[3] https://www.esonetyellowpages.com/
[4] http://vocab.nerc.ac.uk/

- *Research infrastructures*: Especially operators of observatories with instruments but also virtual research infrastructures. Simplified access to shared research infrastructure instrument pools is a key requirement for national, European and multinational infrastructures. Among other benefits, PIDs will facilitate inter-infrastructure loan procedures.

Key impacts

- Enable a global registry of instruments following established practices and infrastructure for the persistent identification of objects.
- Specification of a minimal metadata schema for the description of instruments to be proposed for adoption by PID infrastructure providers, data centers, manufacturers, etc.
- Enable the unambiguous reference of instruments.
- Enable the cross-linking of instruments with data, articles, people and other entities.
- Enable the description of compound and component instruments.
- Enable reference to persistently identified instruments in scientific workflows as well as in provenance metadata, and metadata more generally.
- Contribute to improving data quality and data fitness for reuse, FAIR data and metadata, and trust in data: Computations typically need multivariate streams of data and information about instruments is important for QC/QA, etc.

Engagement with existing work in the area

Persistent identifiers have a long history [2] and persistent identification predates the modern approaches of actionable identifiers resolvable on the Web. Well known for the identification of published literature, PIDs have been applied to an increasing number of entities, objects and agents, digital and physical.

Increasingly popular is the application of DOI to persistently identify published data. Data centers such as PANGAEA have long been collaborating with PID infrastructure providers, in particular DataCite, to implement persistent identification of published data in practice, in workflows researchers use.

More recently, the concept was also applied to physical objects, such as samples. A prominent example is the International Geo Sample Number (IGSN). In addition, physical agents, such as persons (ORCID) and organizations (ISNI) are also being identified. Much progress is being made in cross-linking such identifiers, e.g. literature to data and data to data (Scholix[5] [3]), as well as literature or data to people, especially with the steady integration of ORCID in publishers and data centers (see for instance [4]).

[5] http://www.scholix.org/

By addressing instruments, PIDINST WG merely applies existing approaches for persistent identification to yet another entity. Still, there is a need to get community agreement on how to tackle the details that are specific to instruments. As such, the PIDINST WG can surely adopt many existing ideas and approaches and adapt them to meet the characteristics and needs of instrumentation.

While PIDINST WG focuses and limits its scope to instrument instances, the work relates to registries for instrument models, such as the ESONET Yellow Pages or the NERC Vocabulary Server. Such infrastructure can identify and describe models (types) and PIDINST WG will evaluate the possibility of including a mechanism to link instrument instances to instrument type descriptions.

PIDINST WG relates to systems such as the Rolling Deck to Repository[6] (R2R), a repository for underway data collected during the deployment of research vessels, where data are linked to instruments as they are deployed on vessels during cruises (research campaigns). Interestingly, R2R utilizes DOI to identify cruises (i.e. deployments) and URI to identify instruments, e.g. the Vaisala WXT520 metstation on Atlantis[7]. Another interesting example is the application of DOI to identify seismic networks following the recommendations by the International Federation of Digital Seismograph Networks. An example is the “Seismic network 5E: MINAS Project (2011/2013)” identified by http://doi.org/10.14470/ab466166. The application of DOI to seismic networks was seen as a way to tackle the problem of citation and proper attribution of seismological datasets. The primary requirement is thus data citation.

The issue of providing attribution to instruments and their generated data has also been tackled by creating journals that specialize on the publication of literature describing instruments. An example for such a journal is the Journal of Large-Scale Research Facilities[8] (JLSRF).

Finally, PIDINST WG relates to technology, including PID infrastructure providers, resolution mechanisms, content negotiation, and metadata technologies such as OGC Sensor Web Enablement[9] (SWE), Semantic Sensor Network ontology[10] (SSN), StationXML[11] and CERIF[12]. The use of PIDs to identify sensor web publications using SWE and SSN was identified as a research priority in a joint EU report [5] from Ocean of Tomorrow projects, such as SenseOCEAN[13]. PIDINST WG will address these technologies for persistent identification of instruments.

---

[6] http://www.rvdata.us/
[7] http://data.rvdata.us/page/device/100566
[8] https://jlsrf.org
[9] http://www.opengeospatial.org/ogc/markets-technologies/swe
[10] https://www.w3.org/TR/vocab-ssn/
[11] http://www.fdsn.org/xml/station/
[12] https://www.eurocris.org/cerif/main-features-cerif
[13] http://www.senseocean.eu/

Plan for engagement

PIDINST WG aims to engage a number of parties, most importantly representatives of PID infrastructure providers and institutional instrument database providers, instrument manufacturers, and relevant RDA groups. Such experts will ideally be directly involved in PIDINST WG. It is primarily the task of the current WG members to actively reach out in their communities and to represent PIDINST WG and its work, e.g. in presentations or direct communication. If not engaged directly in PIDINST WG, the WG will maintain bilateral discussions with these parties.

A key community for PIDINST WG to engage with are PID infrastructure providers, such as DataCite, ePIC, IGSN and others. Some of these providers are already represented in PIDINST WG (e.g. ePIC). Current PIDINST WG members have good personal connections to PID infrastructure provider representatives and we plan to engage others actively early on in the WG's development.

A second key community for PIDINST WG to engage with are institutional instrument database providers. Such stakeholders are ultimately the managers of metadata about instruments and the maintainers of landing pages providing such information, i.e. the resources onto which PIDs identifying instruments will be resolved.

PIDINST WG aims at engaging selected manufacturers. This is arguably a difficult task but it may be sufficient to engage a few in order to clarify the possible role manufacturers can play in persistent identification of instruments (e.g. register instruments, maintain landing pages, etc.). Some PIDINST WG members have contacts to manufacturers, or are manufacturers of some of their own instruments. We are thus confident that we will be able to include the perspective of manufacturers in PIDINST WG.

PIDINST WG plans to organize a WG session at each RDA Plenary during the WG's lifetime. These will be the main face-to-face meetings in which to provide an update, make decisions, and discuss the actions over the six-months period. PIDINST WG will also present its progress in the RDA PID IG session at plenaries. Furthermore, PIDINST WG will engage with other relevant RDA WGs and IGs, such as Small Unmanned Aircraft Systems' Data IG or Collections WG, in order to discover and explore synergies. Such engagement is primarily done via joint sessions at plenaries and joining conference calls. Individuals may also utilize regular conferences to organize informal meetings.

Much of the work on engaging these communities will be critically important for the adoption of PIDINST WG deliverables.

Work Plan

Deliverables

Firstly, the PIDINST WG will deliver a white paper on persistent identification of instruments aimed at PID infrastructure providers. This document is a recommendation. The white paper informs PID infrastructure providers in their decision to implement persistent identification of instrument and supports them in the implementation. The document is a result of the work done by PIDINST WG and presents the information collected and developed by the WG during its lifetime. In particular, it will report on cross-community use cases and formulate requirements for PID infrastructure providers based on these use cases.

An important component of the white paper is the proposed schema for metadata about instruments, which PID infrastructure providers may want to consider for instrument registration. This schema will be developed in a suitable format, e.g. XML, and will be available online (e.g. GitHub) as well as documented in the white paper.

Secondly, the PIDINST WG will deliver a white paper on its deliberations regarding the publishing of metadata about instruments by institutional instrument database providers. This document is a technical report. It may cover a wide range of topics, including metadata schema, linking data, content negotiation.

Milestones

Assuming timely formal endorsement by the RDA, the PIDINST WG will hold its kick-off meeting at P11 in Berlin, March 21-23, 2018. Each following Plenary during the WG's lifetime (M1-M18) is considered to be a milestone since the meeting will be an opportunity to (1) update the community on the WG's progress; (2) address and resolve open issues; and (3) plan the actions to be taken over the following six-months period.

The PIDINST WG will reach a milestone with the engagement of one or more manufacturers. Manufacturers should be involved into the discussions and development, either directly at physical or virtual meetings, or otherwise in bilateral communication. The PIDINST WG will reach another milestone with the engagement of one or more PID infrastructure providers interested in adopting the WG's results. Similarly to manufacturers, PID infrastructure providers are preferably involved directly in discussions and development but will otherwise be involved via bilateral communication. Both stakeholders should be involved as early as possible in the WG's lifetime.

Primarily during the first three months (M1-M3), and extending at least into M9, the WG will collect use cases through engaging relevant stakeholders. Two use cases are already available and a couple more are planned. The WG will discuss and analyse these use cases in order to

draw requirements, primarily for PID infrastructure providers. At the end of M3, the WG will have a description of the requirements.

During M1-M6, the WG will develop and present for community review a first metadata schema. A first draft should be published ahead of M6 in order to allow for review by the next Plenary (P12). The metadata schema for PID infrastructure providers will be presented at the P12 meeting (M6).

In addition to revision and continued improvements to the metadata schema for PID infrastructure providers, during M6-M12 the WG will focus on publishing metadata about instruments by institutional instrument database providers. It will further engage stakeholders, in particular PID infrastructure providers, manufacturers, and institutional instrument database providers to catalyse (1) adoption of the proposed schema and (2) gain further requirements for institutional systems that describe instrument instances. At the P13 meeting (M12), the WG will provide an update on the metadata schema (changes) and present the results on its deliberations regarding publishing metadata about instruments by institutional instrument database providers.

During M12-M18, the WG will continue its work to catalyse adoption of the proposed schema with PID infrastructure providers as well as complete and publish the deliverables, specifically the recommendation to PID infrastructure providers and the technical report for institutional instrument database providers. The WG will have a final meeting at P14 (M18). Figure 1 provides an overview of the major milestones and deliverables.

Operation

The PIDINST WG will meet physically at RDA Plenary meetings (P11, P12, P13) during its 18 months lifetime. These meetings will serve as a platform to present and discuss results, address and resolve open issues, and to plan the following six-months phase.

Depending on the size of the community attending conference calls, the WG will organize monthly conference calls for the whole group as well as monthly conference calls for a technical subgroup[14]. The conference calls will be organized every two weeks on a day and time fixed during the kick-off meeting. The group conference calls are intended to update the community and obtain feedback. The technical subgroup conference calls are intended to drive forward the work.

Consensus and conflicts

PIDINST WG plans to develop consensus by encouraging and ensuring participation in meetings, both virtual and physical meetings. PIDINST WG encourages offline thinking and work ahead of conference calls and active discussion during calls in order to draw from a wide

[14] If there is no substantial difference in members between the technical subgroup and the whole group, the WG will organize monthly conference calls and additional intermediate calls on demand.

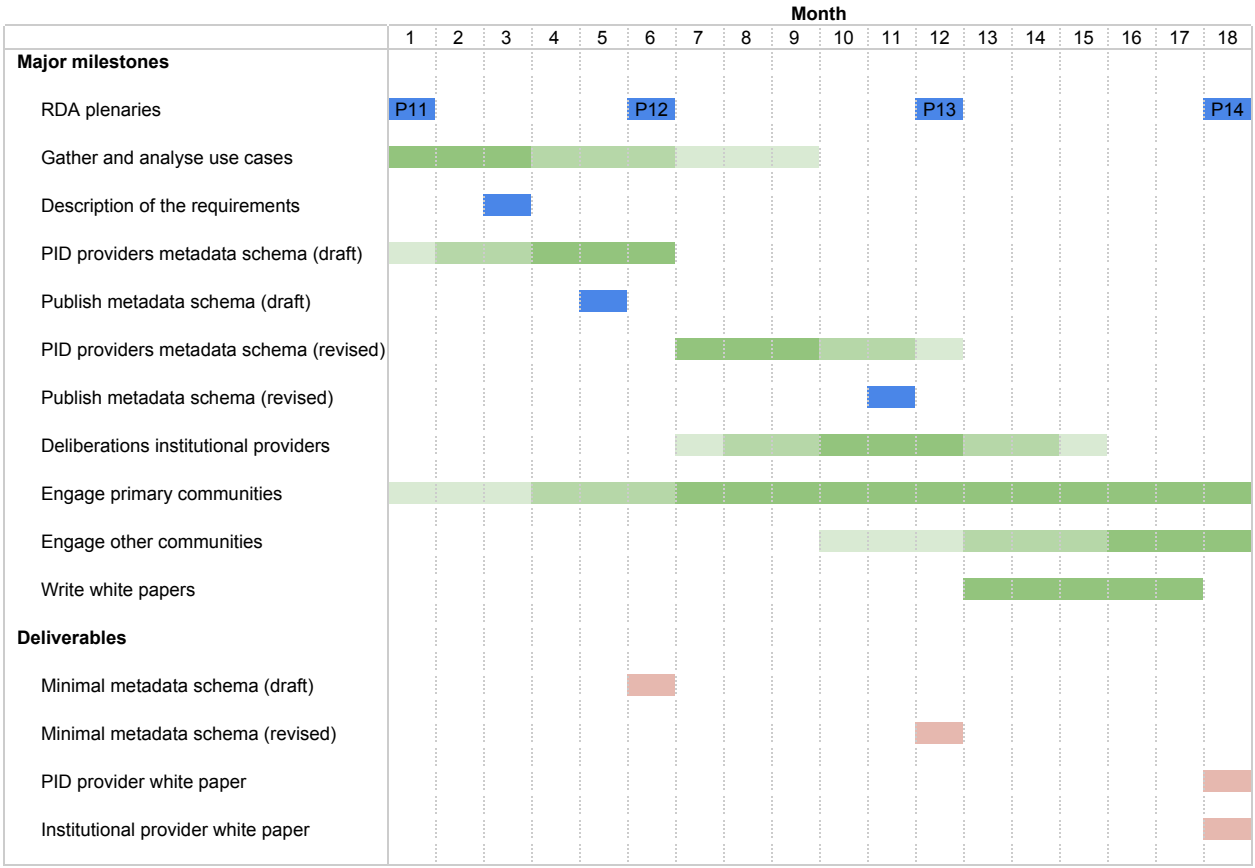

Figure 1: Overview of PIDINST WG Work Plan (M1-18). Primary communities include PID infrastructure providers, institutional instrument database providers, manufacturers.

range of ideas. This should ensure that team members are heard and acknowledged and that the development is not biased toward the thinking of a few members.

If required, PIDINST WG will vote on issues. The WG will adopt a weighted multi-vote approach whereby each member can cast up to three votes, from most to least preferred. This approach ensures members think thoroughly about the available options and decreases chances of stalemate.

The co-chairs are committed to keep the development on track and within scope. The WG will adopt a framework for ranking priorities that helps the team deciding which option will best qualify priorities. We will also undertake regular retrospectives (such as after completing a milestone) where we assess our performance and feedback improvements into the next phase.

Community engagement

The PIDINST WG will engage in particular the following broader communities: Research infrastructures that operate observatories with instruments, data centers, other relevant RDA WGs or IGs, as well as relevant international projects.

National, pan-European, and international research infrastructures that operate instruments are an important and interesting community to engage in this work. Many current PIDINST WG members are either members of one or more research infrastructures or have personal contacts to people involved in such infrastructures. We thus think that the community of research infrastructures will be engaged appropriately. One challenge will be to balance the disciplines (Earth and Environmental Science, Life Sciences, High Energy Physics, Material Sciences, Astronomy, etc.).

As holders of datasets resulting from operating instruments, data centers are another community PIDINST WG should engage with. Of particular interest is the possibility to include PIDs for instruments in metadata of published data as well as cross-linking, i.e. maintaining explicit relations between PIDs for published data and PIDs for instruments. Such relationships are also interesting to maintain provenance.

The PIDINST WG relates to various other RDA WGs or IGs, in particular Collections WG, Data Type Registries WG, PID Kernel Information WG. Since compound instruments and their components can be expressed by a collection of PID references, PIDINST WG naturally relates to the Collections WG. As metadata of instruments should be described in an interoperable and machine readable way, PIDINST WG also relates to the Data Type Registries WG. In some cases, parts of the metadata might be of interest for fast decision processes inside workflows overlooking a huge number of instruments. Such kind of metadata may be described as kernel information inside the PID, which underscores the relationship to the PID Kernel Information WG. Other RDA groups are relevant because they deal with hardware, including instruments. An example is the Small Unmanned Aircraft Systems' Data IG. PIDINST WG aims at engaging other relevant RDA groups through joint sessions at Plenaries or participation in conference calls, e.g. webinars.

Finally, it will be important for PIDINST WG to engage with international projects in the space of persistent identification, such as the H2020 project FREYA; European infrastructure such as the EOSC; as well as international infrastructures, especially those managing instrument databases.

Adoption Plan

PIDINST WG aims at the adoption of its deliverables primarily in two communities: PID infrastructure provider and institutional instrument database providers. Of these two, adoption within the PID infrastructure provider community is the most important.

The white paper for PID infrastructure providers on persistent identification of instruments is the primary PIDINST WG deliverable. PIDINST WG will actively engage the PID infrastructure provider community in a continuous manner in order to communicate development and obtain feedback. Ultimately PIDINST WG envisions that its results will serve as a base to implement persistent identification of instruments in existing PID infrastructure. PIDINST WG aims at avoiding building new infrastructure. Rather, the WG will suggest to build on existing ones.

The second most important community PIDINST WG aims at concrete adoption are institutional instrument database providers. Their systems provide landing pages that describe concrete instruments and are thus key in the resolution of the PID on the web. PIDINST WG not only aims for such providers to adopt the instrument registration mechanism but also at harmonization of the protocols through which human and computer agents can access detailed metadata about instruments as well as harmonization of the presented metadata themselves. Even though such aim is secondary for PIDINST WG, it is clear to the WG that harmonization and support for content negotiation at this level is an interesting objective to keep in mind, possibly for a future WG.

PIDINST WG also aims at adoption among instrument developers and manufacturers. The possibility for such actors to register instruments “at birth” is appealing. Adoption in this space would provide valuable information on feasibility and implications to responsibilities and workflows. Adoption among such actors is planned through active collaboration within the WG in conference calls and physical meetings. While such adoption is interesting, PIDINST WG is conscious of the limited time and scope, and will prioritize adoption in communities of primary interest.

Initial Membership

First name	**Last name**	**Email**	**Affiliation**
Louise	Darroch (co-chair)	louise.darroch@bodc.ac.uk	British Oceanographic Data Centre, National Oceanography Centre (BODC, NOC)
Markus	Stocker (co-chair)	markus.stocker@tib.eu mstocker@marum.de	German National Library of Science and Technology (TIB) and MARUM Center for

			Marine Environmental Sciences (PANGAEA)
Andrew	Janke (co-chair)	andrew.janke@uq.edu.au	NIF Informatics Fellow
Alessandro	Oggioni (co-chair)	oggioni.a@irea.cnr.it	CNR - IREA Milan
Cristiano	Fugazza	fugazza.c@irea.cnr.it	CNR - IREA Milan
Javier	Quinteros	javier@gfz-potsdam.de	GEOFON / Deutsches GFZ Potsdam
Rolf	Krahl	rolf.krahl@helmholtz-berlin.de	Helmholtz-Zentrum Berlin
Ulrich	Schwardmann	uschwar1@gwdg.de	Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Juan	Bicarregui	juan.bicarregui@stfc.ac.uk	STFC
Maggie	Hellström	margareta.hellstrom@nateko.lu.se	Lund University and ICOS Carbon Portal
Alex	Vermeulen	alex.vermeulen@icos-ri.eu	ICOS ERIC-Carbon Portal
Robert	Huber	rhuber@uni-bremen.de	Universität Bremen

References

[1] Borgman, C.L. (2015). Big Data, Little Data, No Data. MIT Press.
[2] Klump, J., Huber, R. (2017). 20 Years of Persistent Identifiers – Which Systems are Here to Stay?. Data Science Journal. 16, p.9. DOI: http://doi.org/10.5334/dsj-2017-009
[3] Burton, A., et al. (2017). The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Magazine, 23(1/2). https://doi.org/10.1045/january2017-burton
[4] de Mello, G., Graef, F., Stocker, M., Schindler, U., Dasler, R., McEntyre, J., Dallmeier-Tiessen, S. (2016). Demonstration of Services to Integrate ORCIDs into Data Records and Database Systems. Zenodo. https://doi.org/10.5281/zenodo.58971
[5] Dunne, D., et al. (2017). Policy Document: Sensor development for the Ocean of Tomorrow. Available at http://www.schema-ocean.eu/Docs/Confirmed/FP7-SCHeMA-614002_Deliverable%20D10.10_v1_29%2009%202017.pdf