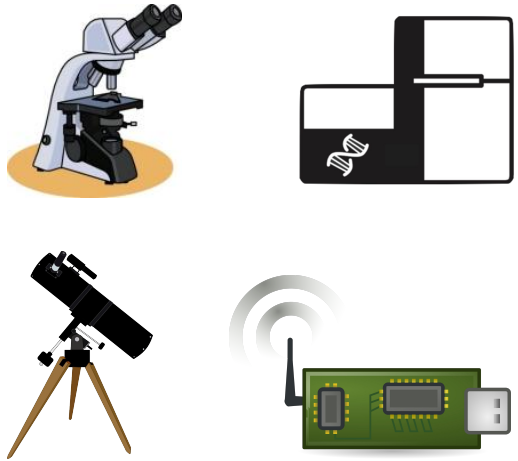
Persistent Identification of Instruments

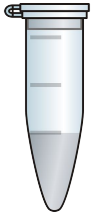
Louise Darroch, Alessandro Oggioni, Cristiano Fugazza, Markus Stocker



bit.ly/2figXYn

Collaborative session notes

PID

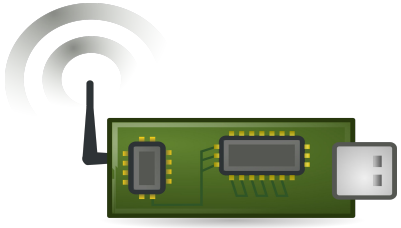


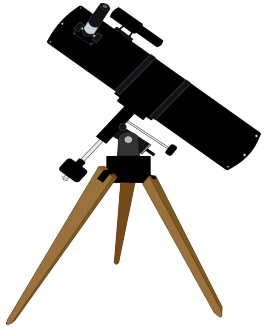
Identification of instruments is not new

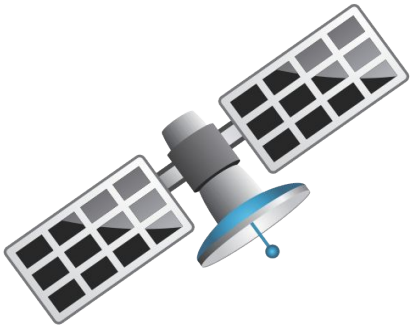
Journal of large-scale research facilities

… articles describing large-scale scientific equipment
… reference large-scale facilities in publications

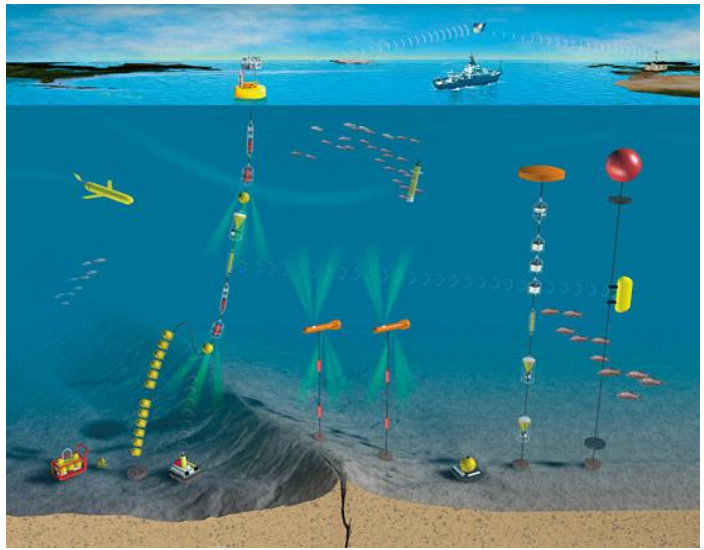
https://jlsrf.org/index.php/lsf

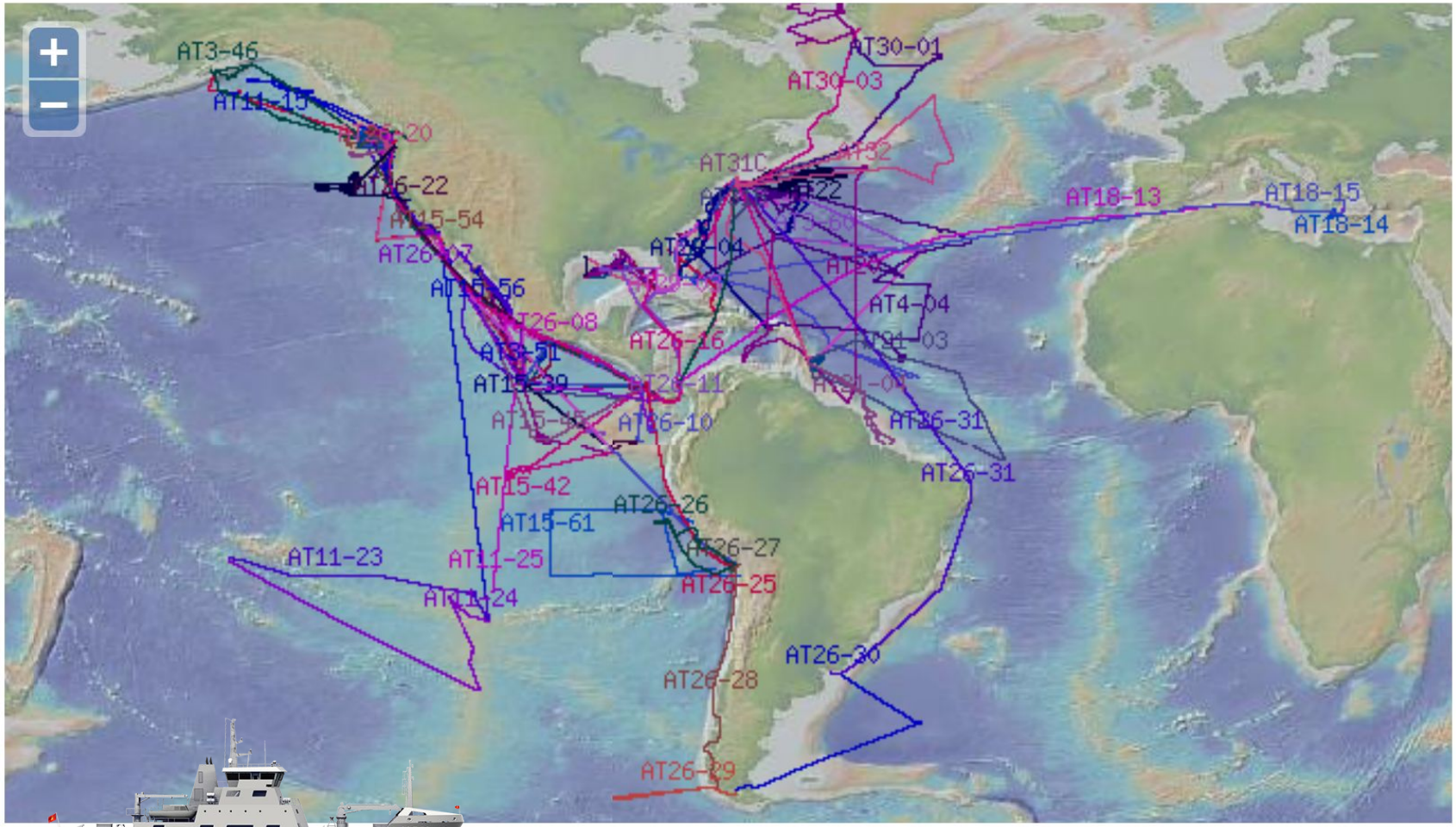







HSV




AT3-46
AT30-01
AT30-03
AT18-13
AT18-15
AT18-14
AT4-04
AT26-31
AT26-26
AT26-27
AT26-25
AT26-30
AT26-28
AT26-29
AT11-23
AT11-25
AT11-24
AT15-42
AT15-61
AT26-10
AT26-16
AT26-08
AT15-56
AT26-22
AT15-54
AT26-07
HSV

“To interpret a digital dataset, much must be known about the hardware used to generate the data, whether sensor networks or laboratory machines.”

“When questions arise [...] about calibration [...], they sometimes have to locate the departed student or postdoctoral fellow most closely involved.”

-- Christine L. Borgman
Big Data, Little Data, No Data
MIT Press, 2015

“To interpret a digital dataset, much must be known about the hardware used to generate the data, whether sensor networks or laboratory machines.”

“When questions arise [...] about calibration [...], they sometimes have to locate the departed student or postdoctoral fellow most closely involved.”

-- Christine L. Borgman
Big Data, Little Data, No Data
MIT Press, 2015

Working Group

- Envisioned is a WG under IG PID umbrella
- Develop a concept for persistent identification of instruments
- Focus on
  - Identifier type
  - Resolution of identifier onto landing pages describing instruments
  - Schema for metadata registration
- Case Statement for P11 Berlin

rd-alliance.org/groups/persistent-identification-instruments

pid-instruments@rda-groups.org

Current state of PIDs for active instruments

LOUISE DARROCH
BRITISH OCEANOGRAPHIC DATA CENTRE (BODC)
NATIONAL OCEANOGRAPHY CENTRE (NOC)

RDA Tenth Plenary Meeting, Montréal, Canada
19th-21st September 2017

Why PIDs?

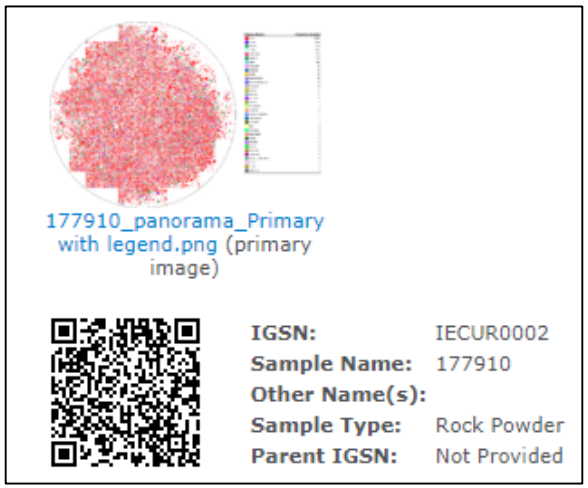
It is customary to think that PIDs are only used to cite journals or datasets….

Classic example: Digital Object Identifier (DOI)

How PIDs are being used

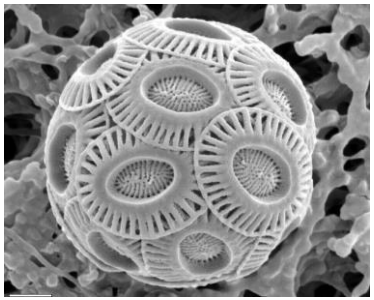
Increasingly, PIDs are being used to universally locate and identify physical things or events

A sample



International Geo Sample Number (IGSN)

A biological entity



WoRMS World Register of Marine Species

Emiliania huxleyi (Lohmann) W.W.Hay & H.P.Mohler, 1967

LSID urn:lsid:marinespecies.org:taxname:115104

Life Science Identifier (LSID)

A researcher

Markus Stocker

ORCID ID

orcid.org/0000-0001-5492-3212

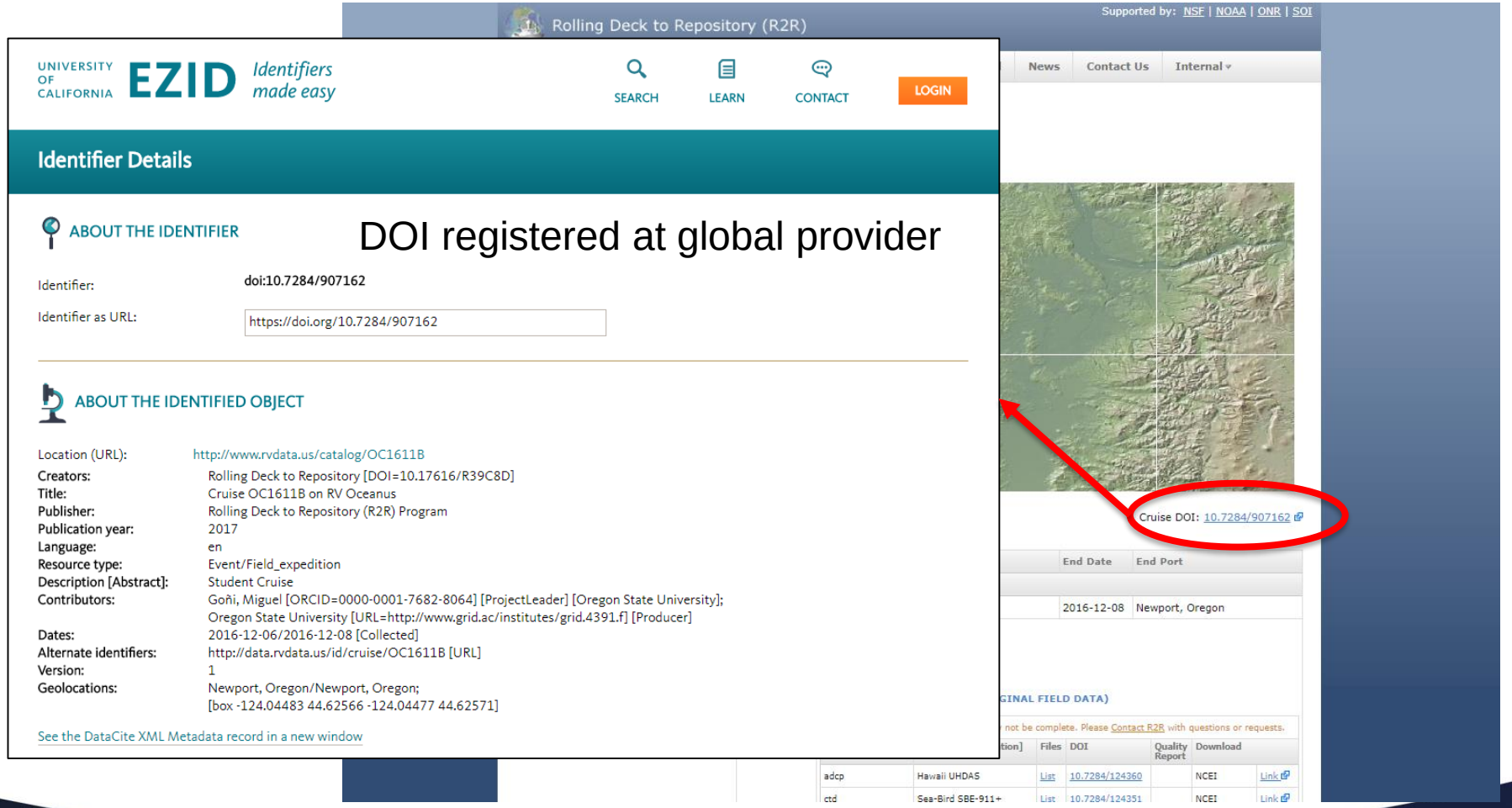
ORCID ID

PIDs and instruments

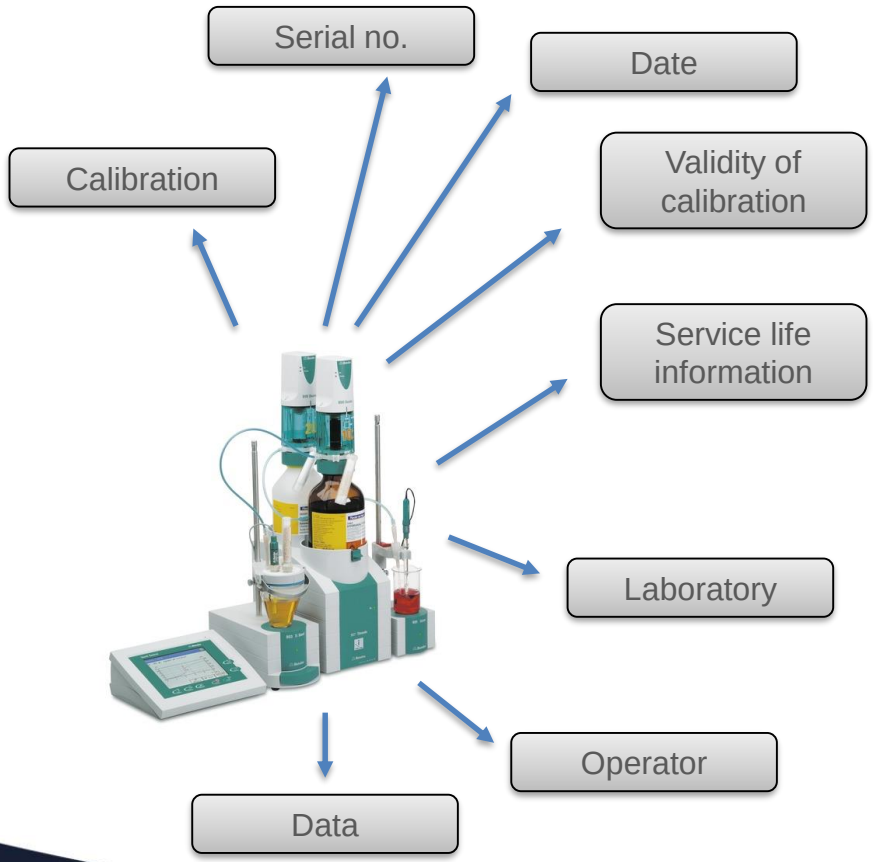
- PIDs are already being used to identify instruments and things related to instruments (some examples below)
- **NOTE:** Not all the same PID types used

What	PID	Thing/event	Who
Platforms	https://doi.org/10.5065/D6DR2SJP	HIAPER Gulfstream GV aircraft	Earth Observing Laboratory (EOL)
Platform instances	http://vocab.nerc.ac.uk/collection/C17/current/32OC/	RV Oceanus	ICES
Deployments	https://doi.org/10.7284/907162	Cruise OC1611B on RV Oceanus	Rolling Deck to Repository (R2R)
Instrument models	SDN:L22::TOOL0882	Rockwell Collins PLGR 96 GPS	SeaDataNet/NERC Vocabulary Server
Instrument instances	http://linkedsystems.uk/system/instance/TOOL0969_1234/current/	Aanderaa 4531 O2 optode (serial #1234)	SenseOCEAN
Data	https://doi.org/10.1594/PANGAEA.879596	Ostracods in permafrost deposits from the Bykovsky Peninsula 1998/1999.	PANGAEA

An example of a deployment

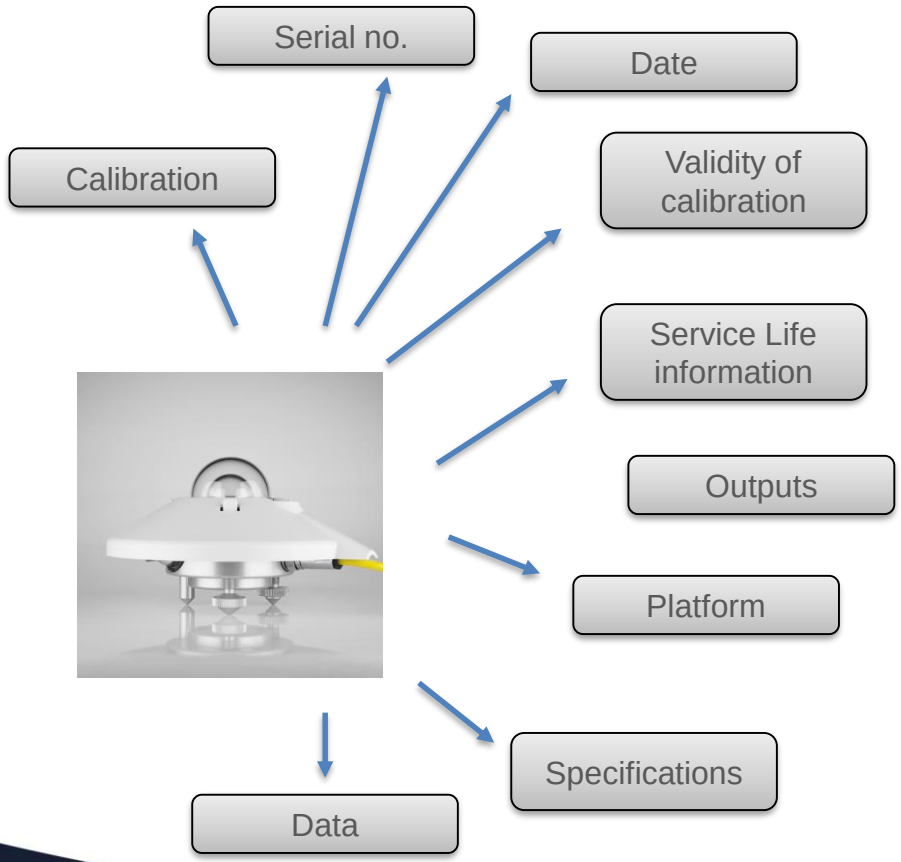


Audit trail



- Linking to the associated metadata about an analytical result is important in some regulated industries (traceability)
- Preventing mix-ups and editing errors gives assurance to data (e.g. climate change studies -> policy)

Audit trail



- Advances in technology mean we are generating more data than ever
- Linking to associated metadata helps us quickly determine if sensors are fit for purpose
- It also enables machines to automate and aggregate sensors and information

What metadata already exists?

- What existing metadata could be resolved under a PID for an instrument instance?

- Many established lists of standardised terms (controlled vocabularies) already in use, especially in the marine domain. E.g.

⇑ -- Hitachi U3010 UV-Visible scanning spectrophotometer --

URI	http://vocab.nerc.ac.uk/collection/L22/current/TOOL0824/
Identifier ()	SDN:L22::TOOL0824
Preferred label (en)	**Hitachi U3010 UV-Visible scanning spectrophotometer**
Alternative label (en)	Hitachi U3010 spectrophotometer
Definition (en)	A bench-top scanning spectrophotometer which operates in the UV-visible wavelegnth range (190 nm - 900 nm). It is fitted with a grating single monochromator (Seya-Namioka mount) which reduces stray light to 0.015%. Sample volumes as small as 5 uL can be analysed.
Version Info ()	1
Deprecated()	false
Broader	http://vocab.nerc.ac.uk/collection/L05/current/LAB20/
Date ()	2015-01-29 11:24:26.0

Example controlled vocabularies

Device type

- SeaDataNet Device Categories (L05)
  (http://vocab.nerc.ac.uk/collection/L05/current/)

Device model

- SeaVox Device Catalogue (L22)
  (http://vocab.nerc.ac.uk/collection/L22/current/)

Outputs

- Climate Forecast Standard Names
- BODC Parameter Usage Terms (P01)
  (http://vocab.nerc.ac.uk/collection/P01/current/)

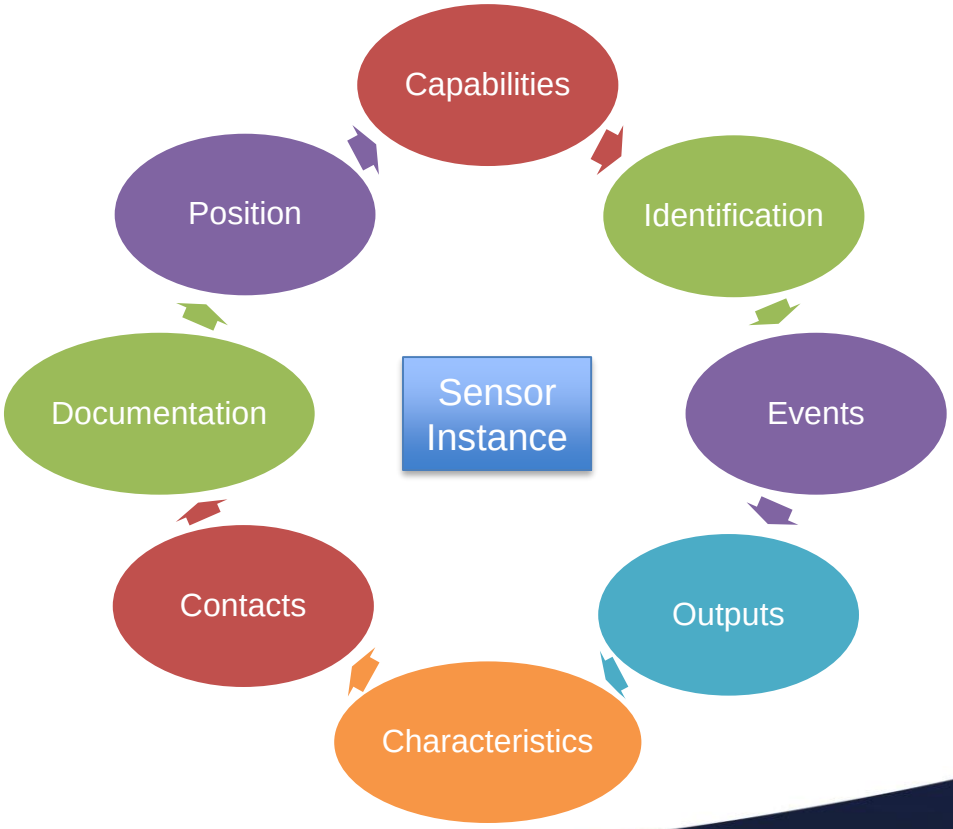
Specifications

- Marine SWE Profiles (W04-W05)
  (e.g. http://vocab.nerc.ac.uk/collection/W04/current/)
- Marine Metadata Interoperability Project Ontology Registry and Repository
  (http://sensorml.com/ont/swe/property)

Individual L22 instrument model published on the NERC Vocabulary Server (NVS2.0)

Instrument metadata schemas

- Schemas have been developed for publishing sensor models and instances on the Semantic Sensor Web
- OGC SensorML
- W3C Semantic Sensor Network



Example of a metadata schema

Open Geospatial Consortium (OGC) SensorML

XML encoding for describing sensors

Enables sensors and processes to be

- better understood by machines
- utilized automatically in complex workflows
- easily shared between intelligent sensor web nodes.

```
<!-- ==================================== -->
<!--           Characteristics            -->
<!-- ==================================== -->
<sml:characteristics name="generalProperties">
    <sml:CharacteristicList>
        <sml:characteristic name="physicalProperties">
            <swe:DataRecord definition="http://sensorml.com/ont/swe/property/PhysicalProperties">
                <swe:label>Physical Properties</swe:label>
                <swe:field name="PhysicalProperties">
                    <swe:DataRecord>
                        <swe:field name="Weight">
                            <swe:Quantity definition="http://sensorml.com/ont/swe/property/Weight">
                                <swe:uom code="oz"/>
                                <swe:value>10</swe:value>
                            </swe:Quantity>
                        </swe:field>
                        <swe:field name="Length">
                            <swe:Quantity definition="http://sensorml.com/ont/swe/property/Length">
                                <swe:uom code="in"/>
                                <swe:value>4.5</swe:value>
                            </swe:Quantity>
                        </swe:field>
                        <swe:field name="Width">
                            <swe:Quantity definition="http://sensorml.com/ont/swe/property/Width">
                                <swe:uom code="in"/>
                                <swe:value>2.5</swe:value>
                            </swe:Quantity>
                        </swe:field>
                        <swe:field name="Height">
                            <swe:Quantity definition="http://sensorml.com/ont/swe/property/Height">
                                <swe:uom code="in"/>
                                <swe:value>1.4</swe:value>
                            </swe:Quantity>
                        </swe:field>
                        <swe:field name="CasingMaterial">
                            <swe:Category definition="http://sensorml.com/ont/swe/property/Material">
                                <swe:value>Aluminum</swe:value>
                            </swe:Category>
                        </swe:field>
                    </swe:DataRecord>
                </swe:field>
            </swe:DataRecord>
        </sml:characteristic>
```

Example of metadata schema

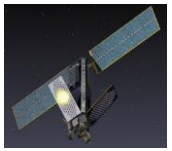
Sensor passes data + UUID through to base station

SensorML & RDF/XML sensor descriptions

Observations & Measurements

Platform

Satellite



SOS, Linked data server

Metadata database and data files

Sense OCEAN

EU Oceans of Tomorrow

- Recently, the SenseOCEAN project used PIDs to locate, resolve and link SensorML (and RDF/XML/SSN) sensor instance descriptions
- They were used to help cut down transmission costs from in-situ sensors
- This was done using a resolvable Universally Unique Identifier (UUID)

http://linkedsystems.uk/system/instance/TOOL0969_1234/current/

Summary

- PIDs are increasingly being used to identify things or events
- Many different PID types are used to identify instruments and things associated to instruments
- There is no universal agreement on one method
- Benefits in linking an active device to associated metadata (e.g. traceability, machine automation)
- Controlled vocabularies to describe metadata associated to sensor instances exist, especially in the marine domain
- Defined metadata schemas are being used for publishing sensor model and instance descriptions on the Semantic Sensor Web

RDA P10: Introduction to ePIC

PIDs for Instruments

Ulrich Schwardmann

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

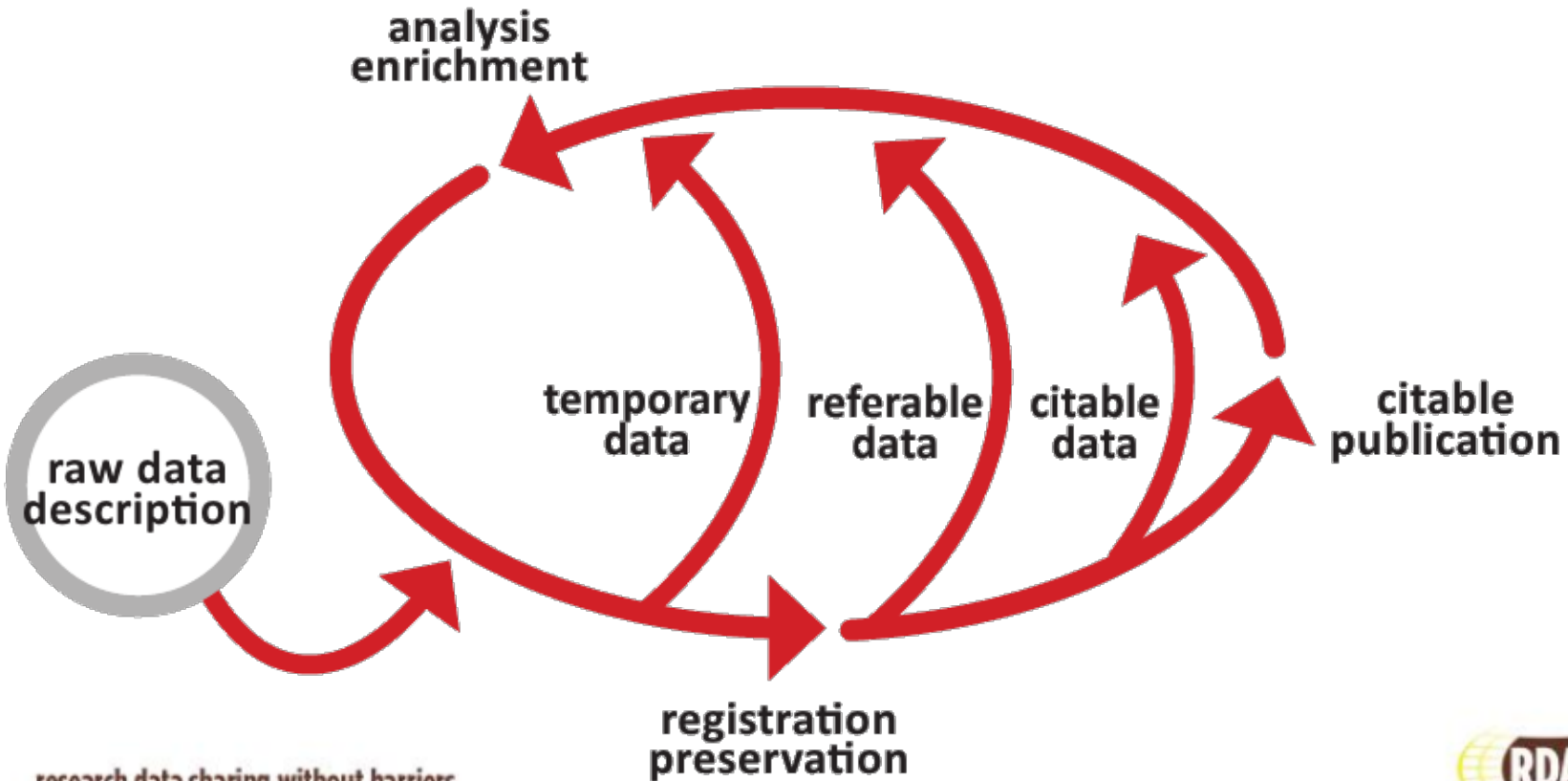
Am Fassberg, 37077 Göttingen
ulrich.schwardmann [at] gwdg.de

21 September 2017, Montreal

The Research Data Life Cycle

data intensive research is highly collaborative

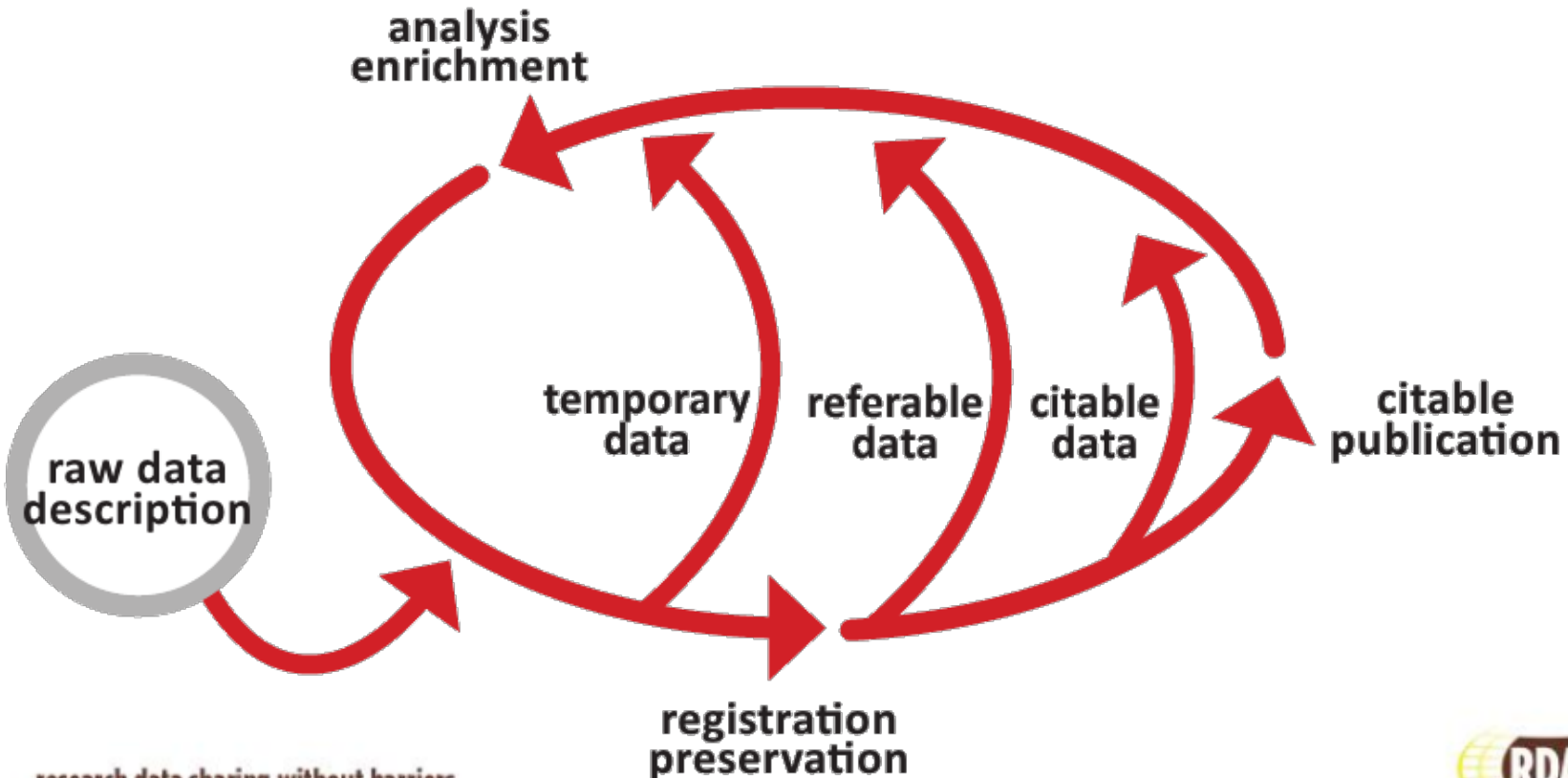
- scientists share data already in an early research state
- ad hoc techniques for sharing are often prohibitive
- reliable references can accellerate the Research Life Cycle



The Research Data Life Cycle

data intensive research is highly collaborative

- scientists share data already in an early research state
- ad hoc techniques for sharing are often prohibitive
- reliable references can accellerate the Research Life Cycle



research data sharing without barriers
rd-alliance.org

The ePIC Members

build a network of currently six strong scientific service providers that signed a contract

- to **ensure a reliable and persistent identifier infrastructure**
- devoted to the needs of the research community at large.

Mayor focus: the referability of data

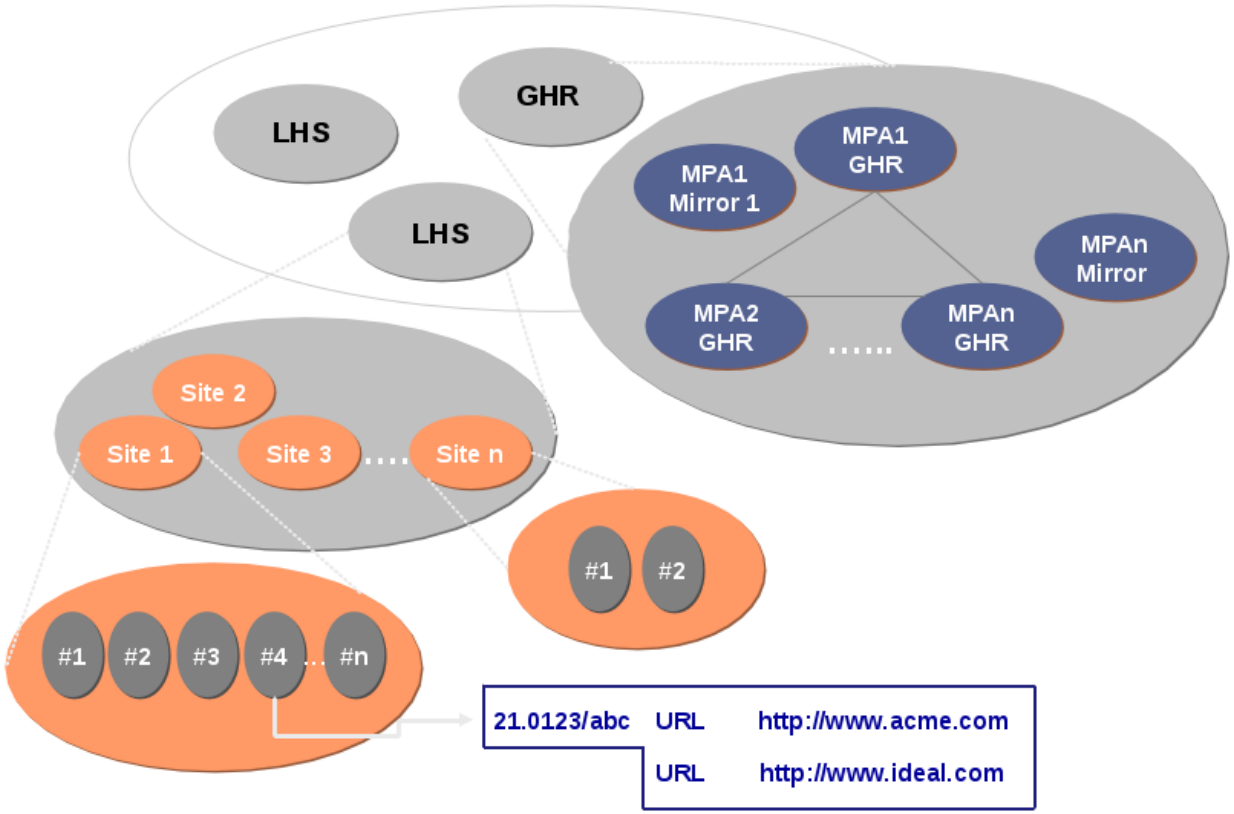
- for sharing during the research process
- with finer granularity and
- PID coupled metadata (PID InfoTypes)

Quality of Service in ePIC

- Conditions of Operation
  - user management, privacy protection and secrecy
- incident management and monitoring
- support system with agreed responsabilties
- certification of ePIC PID services
- several policies for PID minting and update agreed
  - others are still under discussion
- quality of resolution
  - audits can be requested
- community dependend policies (on prefix level)

DONA Handle.Net Multi Primary Administrators

Multi Primary Administrator GHR (since 8th Sep. 2015)



Sharing Data in Research

- data sharing of early results requires
  - a reliable framework of trust
  - transparent and standardized policies
  - registration for referable data
    - stable references
  - **strong coupling between data and metadata**

PIDs can be the pivot to fulfil these requirements

PID Information Types

- are additional metadata stored in the PID database
- intended to be directly accessible independent of any redirection
- typical cases are
  - checksum
  - mime type (incl. version)
  - embargo time
  - expiration date
  - add. metadata file
  - basic Dublin Core
  - access methods, data formats

Granularity

digital objects shared with other scientists for investigation often have a *finer granularity*

- use cases are
  - single experiments
  - simulation output and/or parameter sets
  - single files, tables, pictures, single scanned pages or video/audio sequences
  - sensor outputs (snapshots, dynamic data)
  - software and software versions
- the minting of a huge number of PIDs can be necessary (and favorable)
- in some cases these sets of digital objects are highly structured
  - and accessible by parameterized services
    - this must be recognizable by data types
  - here also *templates* or *fragment identifiers* can be a solution

Templates or Fragment Identifier

rules for strings appended to the PID (see IETF RFC 6570)

- often used to address *service functions* operating on digital objects
- the template implementation in the handle system is simply a rewrite rule
- delimiter and replacement is configurable at prefix level
- *example*
  - delimiter is `@`, which is replaced by `?`
    `11858/00-ZZZZ-0000-0001-CCD1-4@aaa=bbb&ccc=ddd`
  - translates into:
    `http://wwwuser.gwdg.de/~tkalman/downloads/formtest.php?aaa=bbb&ccc=ddd`
- be **careful**: fragment identifier are much **less persistent** then the PIDs itself
- the rewrite rule can be much more complex:
  - replace semantic string elements like URLs by other strings
  - use delimiter strings instead of characters
  - ...

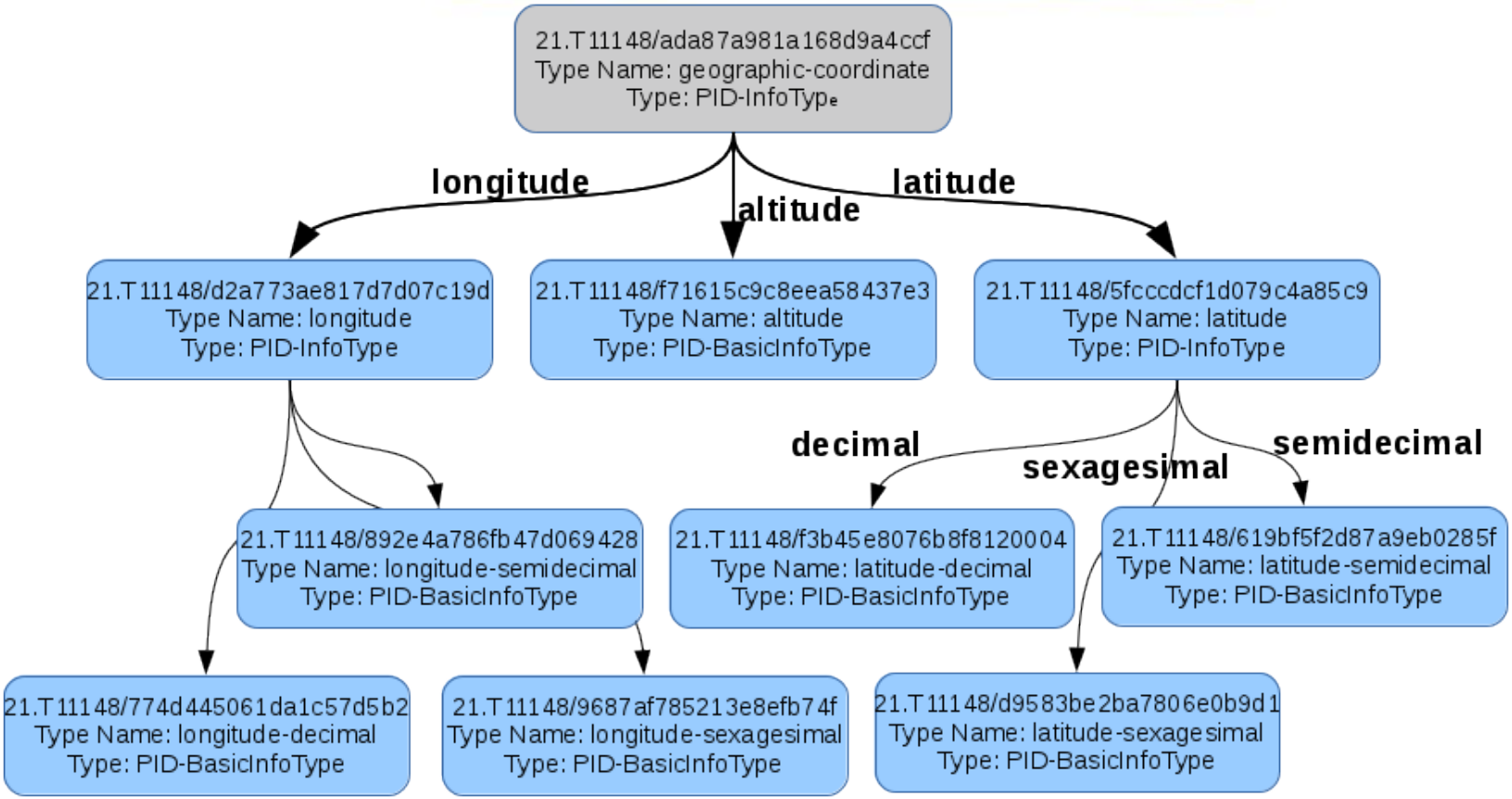
Data Type Registries

- The PID Information Type (PIT) definitions are kept in Data Type Registries (DTRs).
- Currently a couple of such DTRs exist,
  - based on a software called *Cordra*[1], developed from a RDA WG outcome,
  - using a special vocabular for type specifications.
  - This vocabular is partly extended for the purpose of the development presented here.
- ePIC also runs such a DTR
- Interoperability: a process of standardisation and federation for DTRs is ongoing.

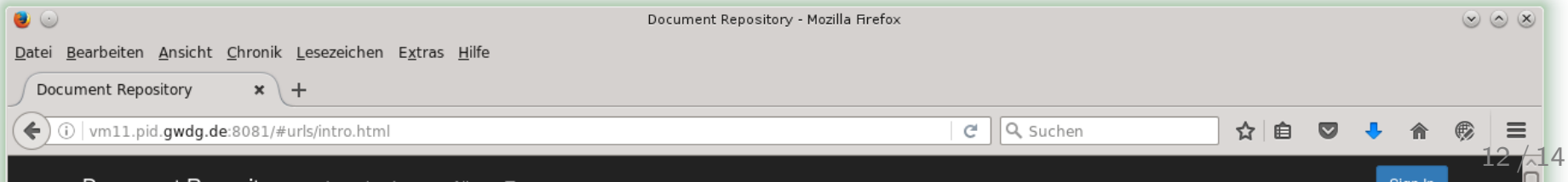
[1] https://www.cordra.org/

Hierarchies in Metadata

Example: geographic coordinate.



- the ePIC DTR can express and validate such hierarchies

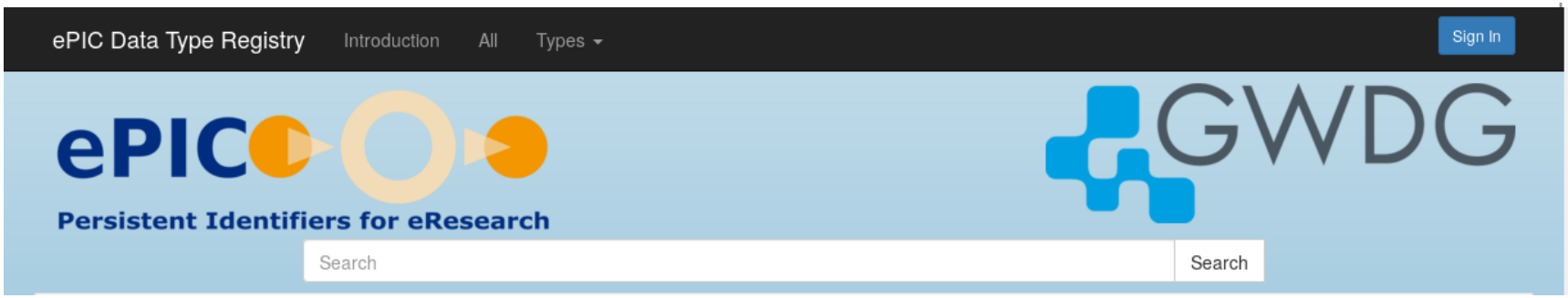


The ePIC DTR Homepage

`http://dtr.pidconsortium.eu/`

PID InfoType states are:

- *in preparation* (21.T11148),
  - `http://dtr-test.pidconsortium.eu/`
- *candidate*, *approved*, *deprecated* (21.11104)
  - `http://dtr-pit.pidconsortium.eu/`



Many Thanks

Questions ???

Contact@ePIC:

- support@pidconsortium.eu

Contact@GWDG:

- **Ulrich Schwardmann**
  T: 0551 201-1542, E: ulrich.schwardmann@gwdg.de

- Focus on
  - Identifier type
  - Resolution of identifier onto landing pages describing instruments
  - Schema for metadata registration
  - Content negotiation and machine readability
- Many projects (in Earth science) building “sensor registries”
  - Can this WG lay the foundations for a global instrument registry?
  - Deliver a recommendation for an organization to implement, run service
- Instruments, great but
  - Also platforms and deployments
  - Links between them
- Involving manufacturers
  - Do we need to involve them
  - Should they register instruments and provide landing pages
- PID type and resolution mechanism: Existing or new?
- Involve disciplines: earth science, astronomy, life sciences, chemistry, ...
- Co-chair from US/Australia