Persistent Identification of Instruments WG (PIDINST WG)

The PIDINST team

tinyurl.com/ybbalyzf

PIDINST WG?

- Information about instruments plays an important role in science
- Sources of data and knowing which instrument and its properties matters

Seeks to propose a community-driven solution for globally unique and unambiguous identification of instrument instances that are operational in the sciences

- Leverage on existing PIDs and PID infrastructure

Potential Benefits

- Link data to the instruments that generated them (provenance)
- Aid equipment logistics and mission planning
- Facilitate interoperability and open data sharing
- Improve the discoverability and visibility of instruments and their data
- Metrics that quantify the use of instruments
- ...

Status Update since P10 (Montreal)

- Delivered the Case Statement
- Obtained TAB Review and RDA endorsement
- Clap, clap, clap ;>
- Collected five use case descriptions
- Regular monthly conference calls

Case Statement Overview

- Key beneficiaries (among others)
  - Researchers: Contextual information to determine how to process data
  - Data repositories: Link to PIDs at instance granularity in metadata
  - Hardware curators: Support keeping track of institution’s instruments
  - Manufacturers: Could play major role in instrument registration and metadata management
- Key impacts
  - Enable a global registry of instruments
  - Specification of metadata schema for instrument description (PID infrastructure)
  - Enable reference to instruments in scientific workflows
  - Contribute to improve data quality and fitness for reuse, FAIR data and metadata, trust in data
- Engagement with
  - Existing work: PIDs, model registries, existing systems and vocabularies
  - Stakeholders: PID infrastructure, instrument databases, manufacturers, relevant RDA groups
- Work plan presented later by Louise

TAB Review

- Positive
  - The objectives and deliverables are well aligned with the RDA mission and the scop
  - Very worthwhile effort
  - If successful will be a very positive contribution associated with RDA
  - Outcomes will be welcomed by the PID community
  - Improve the precision of data sharing and interpretation
- Suggestions
  - Greater variety of disciplines, instrument types
  - Potential uptake by manufacturers
  - Engage academia and industry (how are cellular phone unique numbers managed globally?)

GEOFON use case

FDSN standard and recommendations

- GEOFON is one of the most advanced seismological data centres in Europe.
- As almost all seismological data centres, GEOFON follows the recommendation from the International Federation of Digital Seismograph Networks (FDSN, 2014).
- According to it, each seismic network must be identified by a DOI and metadata has to include at least certain fields related to the Datacite format.
  - Creator, Title, Publisher and Publ. year are mandatory.
  - Resource type, Description and Format are recommended.
  - Contributor, Location, Size, Date Collected, Date available and Relat. IDs are optional.

FDSN standard and recommendations

From the recommendations it can be seen that there is a fuzzy line which separates the hardware, the metadata describing it, and the data.

“In this view a seismic network is an entire collection of sensor data, but also the seismic metadata associated with it, such as station details, instrument types, response data.”

Other needs for the Pool Management

- Our Pool Management Team (GIPP) has also the need to keep track of all the hardware components, different deployments and calibrations.
- In particular, technical specifications of deployed stations, identifying particular instances of the sensors and *not only the type/model*.
- a journal of the different components could be offered through landing pages.
  - where has been used? for how long?
  - were there problems with it? how have they been solved?
  - has it been recalibrated?
- Provenance data from all these points.

Other needs for the Pool Management

- Then, GEOFON could link datasets with hardware components and inventory metadata.
- Also, provenance data generated could be linked to datasets, offering the user more elements to evaluate the quality of the data.
- Information on which stations were built during field trips could be extremely useful for early detection of problems and to find solutions.
- In an ideal case, new deployments can be informed online, keeping a “live” view of the campaign.

HZB Helmholtz Zentrum Berlin

Use Case for Photon and Neutron Facility

Rolf Krahl

Persistent Identification of Instruments WG @ RDA P11,
Berlin, 21 March 2018

HZB

- Helmholtz Zentrum Berlin für Materialien und Energie (HZB) operates synchrotron light source BESSY II.
- Extremely brilliant synchrotron light pulses with adjustable wavelength, polarization, and photon energy are used as probe to examine various kinds of samples.
- More than forty experiment stations, large variety of methods and experimental techniques.
- Experiment stations may either be fixedly attached to a respective beamline or flexible and can be moved between beamlines.

Instruments at Photon and Neutron Facilities

Particularities of instruments at PaN facilities:

- Multiple complex instruments involved in a single measurement: source, insertation device, beamline, experimental station.
  $\Rightarrow$ *May need to reference a combination of instruments at once.*
- Unique instruments. Mostly designed and sometimes even manufactured in-house.
  $\Rightarrow$ *There may be no external manufacturer, no standard type.*
- Built off several components: simple (mirror, slit), complex custom built (monchromator), off-the-shelf products (detectors).
  $\Rightarrow$ *May need to also identify individual components.*
- Setup may change over time.
  $\Rightarrow$ *Need some kind of versioning.*

Use cases and benefits

- Document the provenance of datasets.
- Track the scientific output of a given instrument.
- For a given dataset, search for other datasets created at the same instrument. Search for calibration data.
- Each HZB instrument has a web page providing documentation on the instrument, its design, and capabilities. Link this page from the PID.
- Attribute PIDs also to major components of an instrument, such as the detector. This allows an independent description of the characteristics of these components.
- Provide relevant metadata that can be automatically retrieved for any objects referencing the PID. E.g. the metadata schema for datasets created by the instrument.

Properties

- Obvious attributes: name, description, manufacturer, type, owner, landing page, . . .
- Reference technical specification.
- Life time: start and end date of the instrument being in operation.
- Documentation: have a “is described by” relation with other resource.
- Versioning: have a “is new version of” and “is previous version of” relation with other instrument.
- Components: have a “has component” and “is component of” relation with other instrument.
- Extensible: Link other related resources.

Journal of Large-Scale Research Facilities (JLSRF)

- Earlier approach to address some of the use cases: JLSRF.
- HZB's instruments have an article in JLSRF describing the instrument.
- Users are asked to cite this article in papers using data created at the instrument.
- The DOI of the JLSRF article is partly used as an substitute for the (not yet existing) instrument PID.
- Nevertheless, both approaches are not redundant: the textual instrument description in JLSRF gives more value to a human reader, while the instrument PID provides much richer options to automatically aggregate information by following the references.

National Imaging Facility

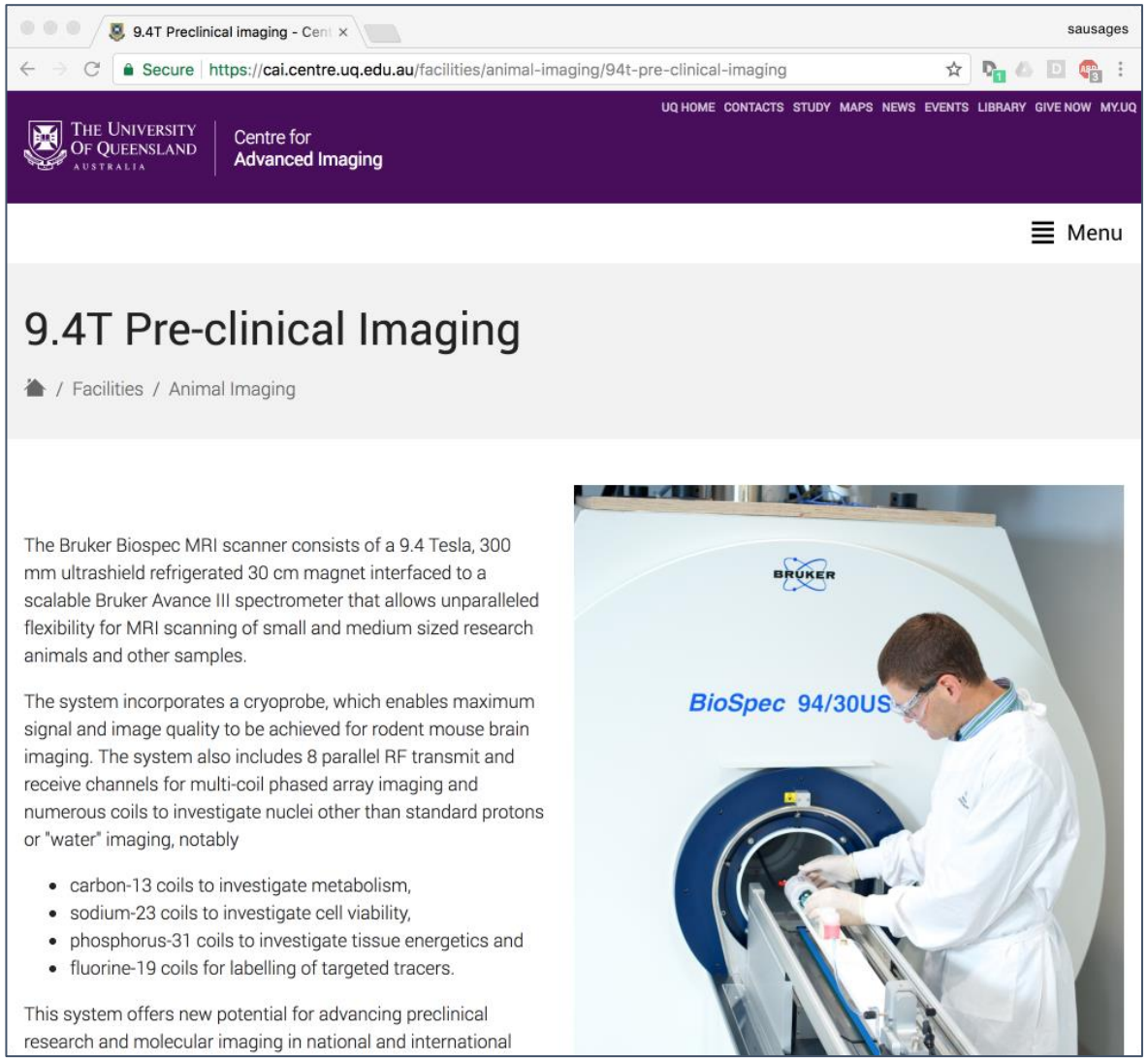
NIF use case

Andrew Janke + Siobhann McCafferty (DLCF)
www.anif.org.au
www.dlc.edu.au

ands nectar RDS www.rds.edu.au aarnet Australia's Academic AUSTRALIAN

NIF needed identification of instruments

- Persistent tracking of provenance of datasets
- Persistent tracking of QC associated with instruments and linked to data
- "9.4T Bruker MRI at the Centre for Advanced Imaging" doesn't work

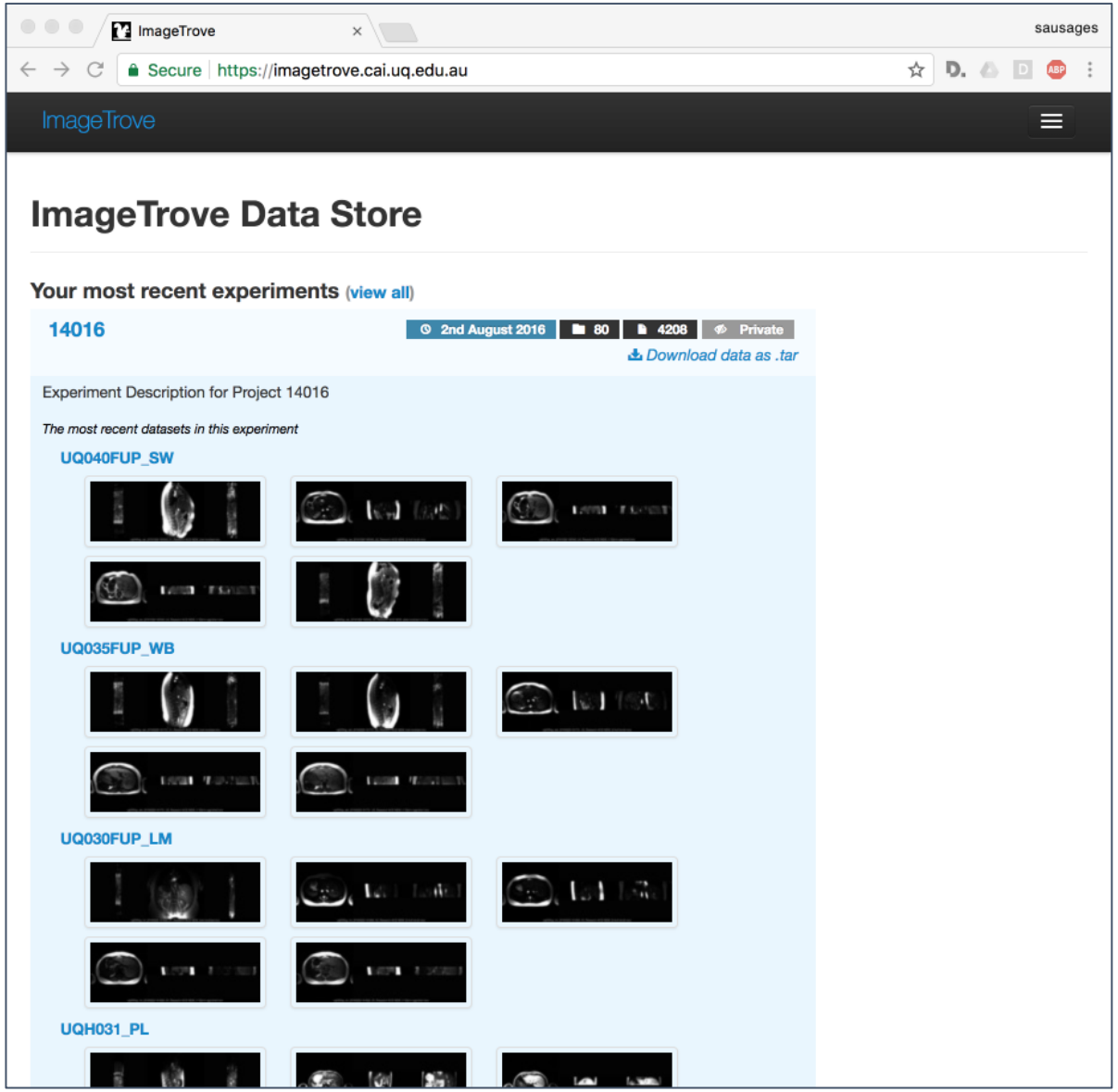


Imagetrove - Data repository

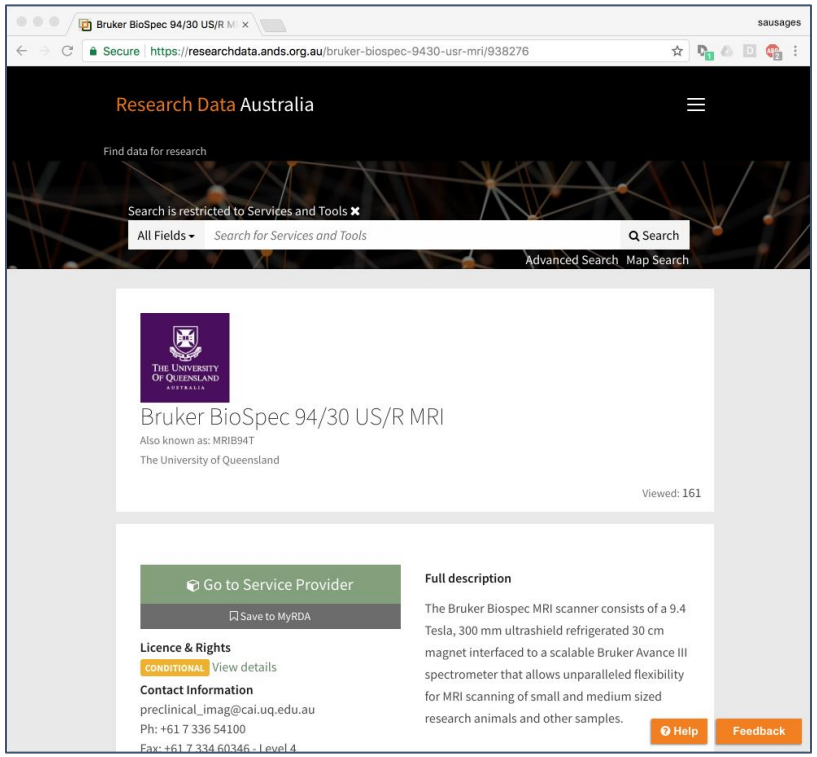
- Data by project (RAiD)
- People by ORCiD

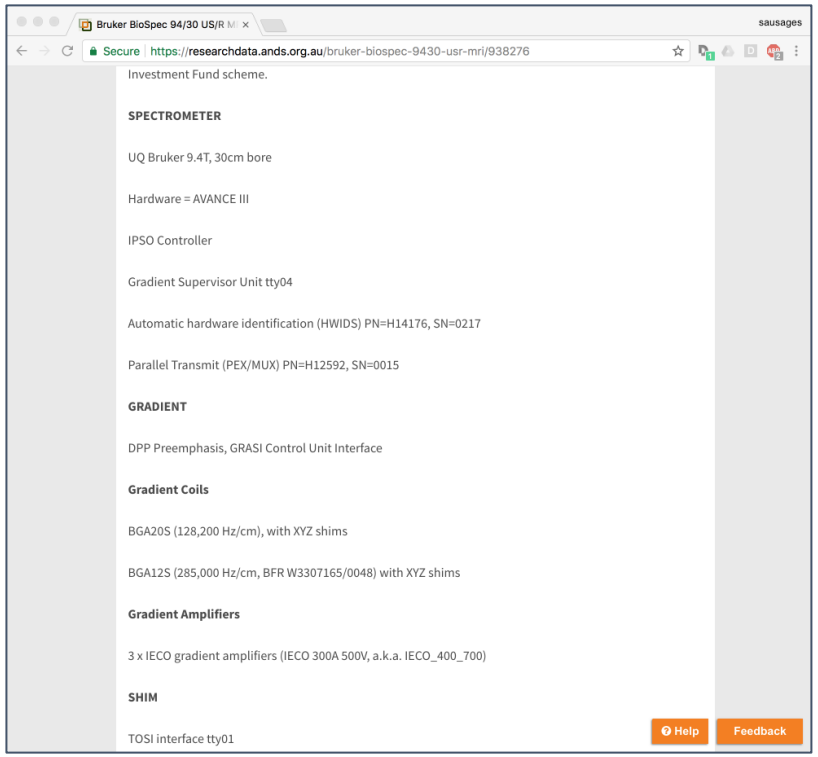
- But data still linked to Instrument Names "9.4T Bruker"

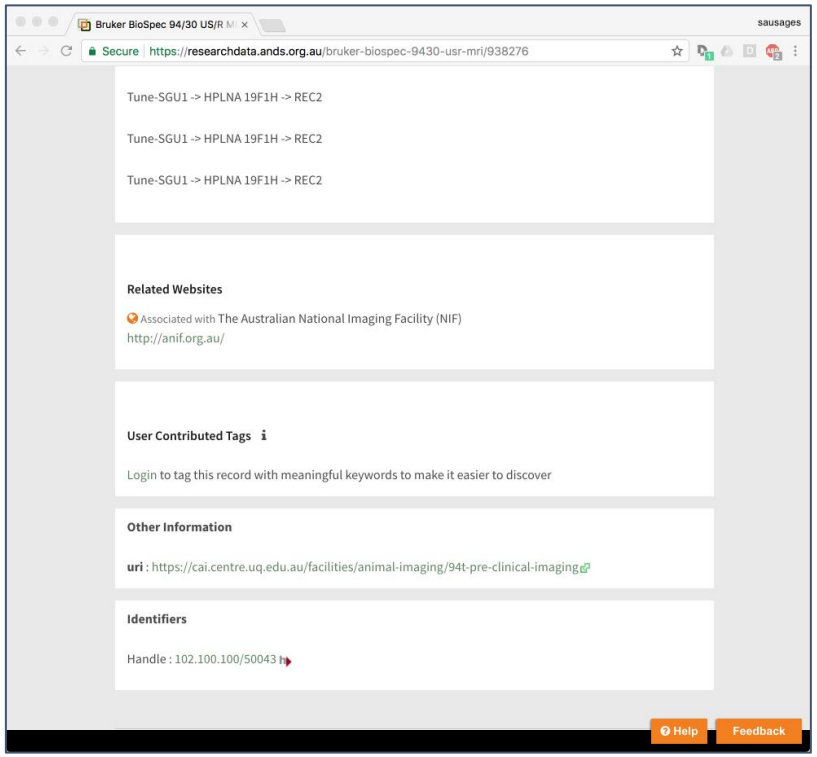
- Use ANDS (Australian National Data Service) Handle service linked to a service record.



ANDS - Service Record - Instrument Config



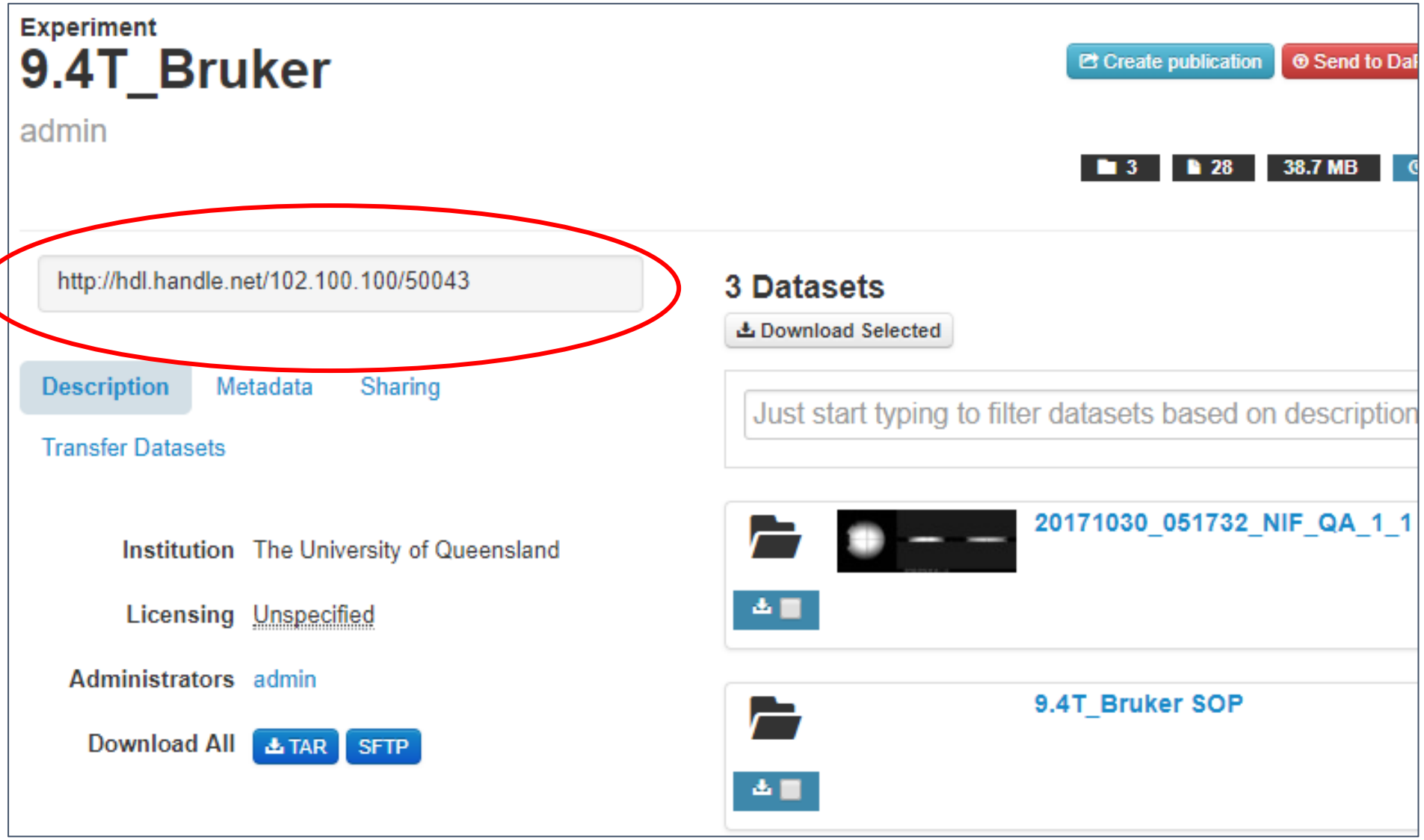




https://researchdata.ands.org.au/bruker-biospec-9430-usr-mri/938276

http://hdl.handle.net/102.100.100/50043

ImageTrove - Link to data



NIF Certified data

National Imaging Facility

NIF agreed process for acquiring trusted data

Contributors	Andrew Mehnert (UWA), Andrew Janke (UQ), Marco Gruwel (UNSW), Wojtek Goscinski (Monash)
Version	1.0
History	5 March 2017 - Draft 25 May 2017 - V1.0 7 December 2017 - minor edits

REQUIREMENTS THAT MUST BE SATISFIED FOR DATA TO BE NIF-CERTIFIED

Each data set acquired on an instrument is deemed to be associated with a Project that has a unique identifier hereinafter called the *Project Identifier* (PID). In the simplest case this will be a site-specific unique identifier. Looking to the future, the Research Activity Identifier Project (RAiD), in development, offers the possibility to mint persistent identifiers for projects.

For data to meet the definition of NIF-certified it must:

1. Have been acquired on a *NIF-compliant instrument*;
2. Possesses *NIF-minimal metadata* including a cross-reference to the relevant instrument QC data;
3. Include the native data generated by the instrument in proprietary format, including the acquisition settings/parameters; and
4. Include conversions to one or more open data formats.

The intent is that these requirements are necessary and sufficient for a user or re-user to establish the provenance and quality of the data and to determine whether it is fit-for-purpose.

NIF-compliant instrument

An instrument is deemed to be NIF-compliant if each of the following is true:

1. It has been assigned a unique identifier, hereinafter called the *Instrument Identifier* (IID). This identifier should be a handle minted through the following ANDS service: http://www.ands.org.au/guides/identify-my-data. The handle should resolve to the RDA record in 2.
2. A record has been be registered for the instrument in Research Data Australia (RDA) that contains the IID as the identifier (resolving to the record itself), a detailed description of the instrument, and related websites. See here for an example: http://hdl.handle.net/102.100.100/50041
3. A documented quality control (QC) process exists for the instrument including definitions of quality assurance (QA) measures, e.g. the signal-to-noise ratio (SNR).
    - For QA measures not provided by the instrument, standardised and reproducible workflows must be defined and made available through a version control repository, e.g., github.
    - Where possible the QC process should be standardised across NIF nodes that possess the same or similar instrument.
    - The QC document must define who has access to the QC data; e.g. access might be restricted to just the instrument/facility manager.

NIF Data Certification	
NIF Certification	Certified
Associated QC Dataset	20171030_051732_NIF_QA_1_1
Instrument record	http://hdl.handle.net/102.100.100/50043
Acquisition Time	2017-11-01 02:00:00+00:00

Issues

- Handle ID's while functional are not granular. A different configuration of an instrument -> new Handle ID and new record
- What NIF is doing is not an international standard -> Thus the certification is only recognised within NIF.
- Adoption of a standard would mean any dataset that is used has a persistent instrument identifier

Sensor Web Enablement (SWE) and Semantic Senor Network (SSN)

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

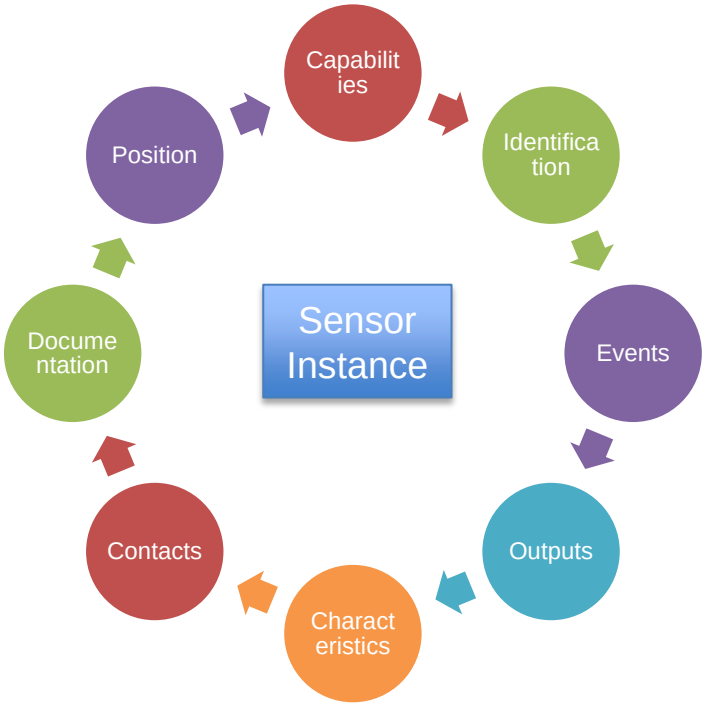
RDA 11th Plenary Meeting, Berlin, Germany
21st-23rd March 2018

What is SWE and SSN?

- Standards and ontologies for making sensors discoverable, accessible and usable via the Web

Semantic Sensor Network (SSN) Ontology

Sensor Web Enablement (SWE)



- Describe sensors - fitness
- Machine readable
- Automate workflows
- Shared between global nodes

Example of SWE (SensorML)

linkedsystems.uk/system/instance/TOOL0969_1234/current

```
<?xml version="1.0"?>
- <sml:PhysicalSystem gml:id="TOOL0969_1234" xsi:schemaLocation="htt
http://www.opengis.net/sensorml/2.0 http://schemas.opengis.net
http://schemas.opengis.net/iso/19139/20070417/gmd/gmd.xsd ht
http://www.opengis.net/gml/3.2 http://schemas.opengis.net/gml/
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:sml="http://ww
xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:xsi="http://
    <gml:description>TOOL0969_1234</gml:description>
    <gml:identifier codeSpace="uniqueID">http://linkedsystems.uk/sy
  - <sml:keywords>
    - <sml:KeywordList>
        <sml:keyword>O2Sat_2</sml:keyword>
        <sml:keyword>Temp</sml:keyword>
        <sml:keyword>WC_dissO2_uncalib_2</sml:keyword>
      </sml:KeywordList>
    </sml:keywords>
  - <sml:identification>
    - <sml:IdentifierList>
      - <sml:identifier>
        - <sml:Term definition="TOOL0969_1234">
            <sml:label>UUID</sml:label>
            <sml:value>TOOL0969_1234</sml:value>
          </sml:Term>
        </sml:identifier>
```

```
<!-- ================================== -->
<!--           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>
```

The use of a unique identifier

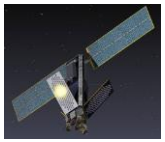
Sensor passes data + UUID through to base station

SensorML & RDF/XML sensor descriptions

Observations & Measurements

Platform

Satellite



SOS, Linked data server

Metadata database and data files

Sense OCEAN

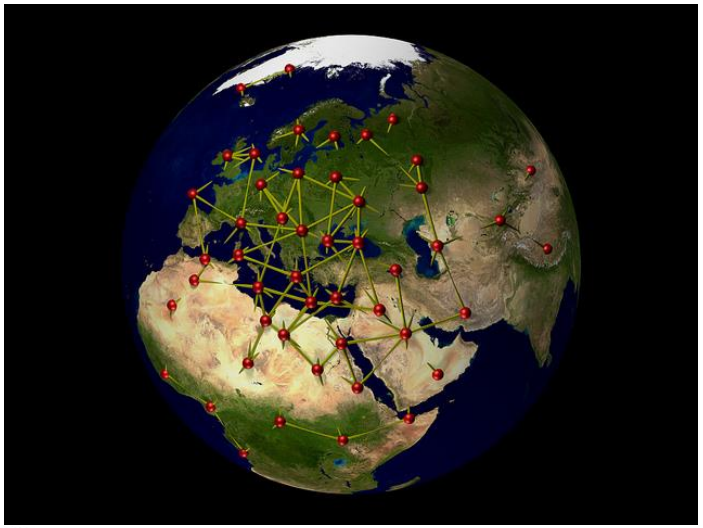
EU Oceans of Tomorrow

- Resolve Sensor Web Publications
- Helped cut down transmission costs
- Used a resolvable Universally Unique Identifier (UUID)

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

A globally unique identifier

Sensors are everywhere



Facilitate sharing between sensor nodes

A global sensor network?

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

Things to consider

- The type of PID

  IPv6 - 2001:0db8:0000:0042:0000:8a2e:0370:7334

  DOI - https://doi.org/10.1109/5.771073

- Publication of standardised metadata schema

```
<!-- ==================================== -->
<!--           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>
```

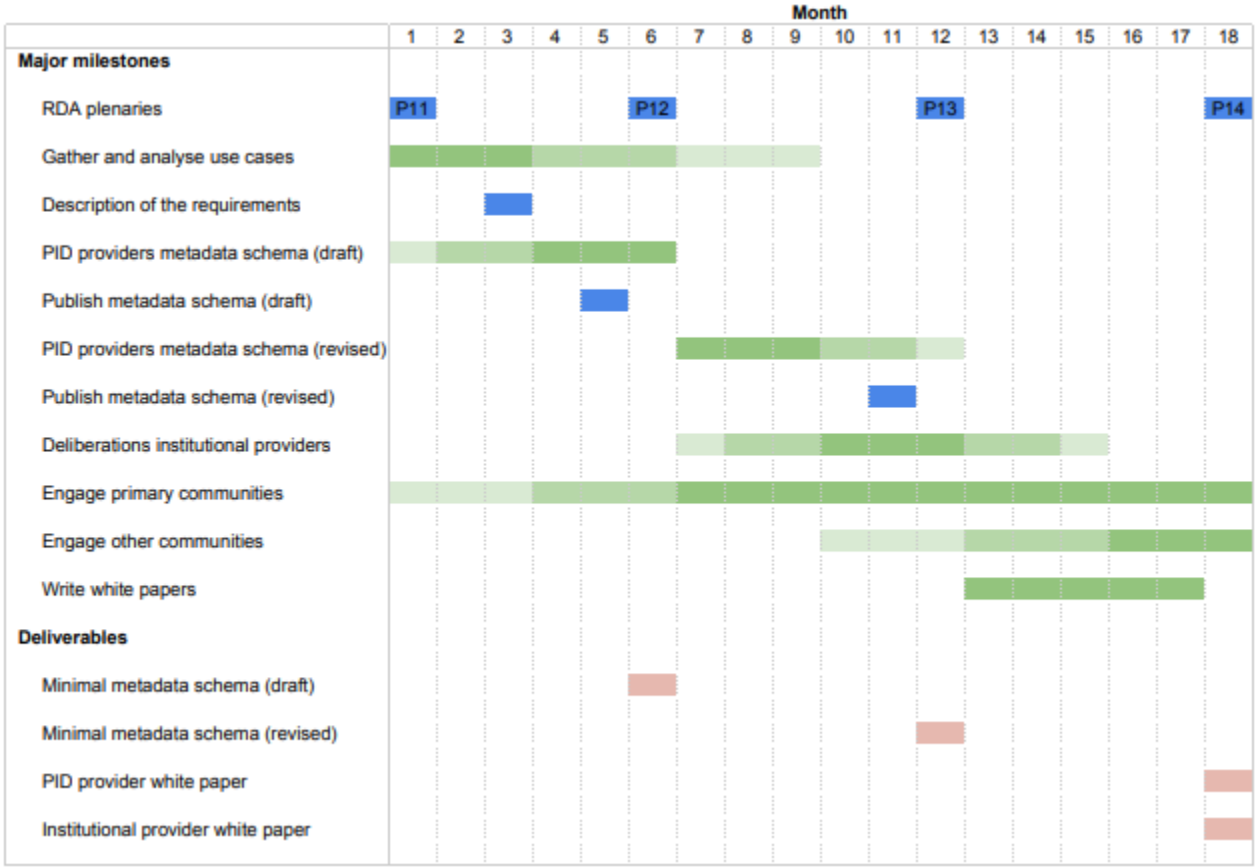
- Handling duplicate sensors, versioning, ownership etc.

Future

- Pursue wider discussions and input
- Liaise with international working group:
  Marine Profiles for OGC SWE Standards Team
- Welcome other input to this use case

Working group overall work plan

https://www.rd-alliance.org/sites/default/files/case_statement/rda-wg-pidinst-case-statement.pdf



Deliverables (by month 18)

1) PID provider white paper (recommendation report)
   - Aimed at PID providers
   - Based on cross-community use-cases
   - Proposed schema for instrument metadata

2) Institutional provider white paper (technical report)
   - Aimed at institutional instrument providers
   - Advise on publishing institutional metadata
   - Wide range of topics (e.g. content negotiation, linking data)

Work plan (first 6 months)

- P11 (Berlin) and P12 (Botswana) plenaries
- Gather and analyse use cases
- Describe the requirements
- **Draft** a metadata schema for PID providers
- Engage primary communities
    - Manufacturers
    - PID providers
    - Institutional instrument database providers

Mode of operation

- Monthly group conference calls
  - Updates
  - Feedback
- Monthly technical sub-group calls
  - Drive forward the work
- Calls are at EU/AU/US friendly times
- Interval and times to be decided.

Discussion

- Definition of instrument
  - Sensor - Device, agent (including humans), or software (simulation) involved in, or implementing, a Procedure. Sensors respond to a Stimulus, e.g., a change in the environment, or Input data composed from the Results of prior Observations, and generate a Result. Sensors can be hosted by Platforms. (Semantic Sensor Network Ontology)
- Supply use cases
  - What other communities should be involved
  - Anyone in the audience with a use case?
- Outreach to include greater variety of disciplines, instrument types
- How to involve manufacturers
  - How to make it appealing to them, e.g. through large infrastructures using their instruments
- Engage PID infrastructures