Design and Implementation of a
Port-Log Discovery Service

Project Report

Submitted to

**MEDIN**

by

**OceanWise Ltd**

November 2019

**Revision and Approval for Use**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Revision Number** | **Revision Date** | **Reason for Revision** | **Author** | **Checked** | **Approved** |
| 0.1 | 2019-10-08 | Draft for comment | MJO | JR | AMJ |
| 0.2 | 2019-11-27 | Pilot metadata successful  |  |  |  |
|  |  |  |  |  |  |

**Distribution**

This document has been distributed to the persons named below and no wider distribution is permitted without the written approval of the document owner.

|  |  |  |
| --- | --- | --- |
| **Name** | **Organisation** | **Title** |
| Clare Postlethwaite | MEDIN |  |
| Sean Gaffney | MEDIN |  |
| Internal | OceanWise |  |

**Copyright**

This document and the contents herein belong to OceanWise Ltd. However, they may be reproduced free of charge in any format or medium provided that the reproduction is accurate and not used in any misleading context or in a derogatory manner. The material must be acknowledged as an OceanWise Ltd publication and the publication cited when being reproduced as part of another publication or service.

**Disclaimer**

The information contained in this publication has been compiled from sources believed to be proven in practice and reliable but no warranty, expressed or implied, is given that the information is complete or accurate nor that it is fit for any particular purpose. All such warranties are expressly disclaimed and excluded, and users are therefore recommended to seek professional advice before committing to investments that incorporate any materials described in this document.

Contents

[1 Introduction 4](#_Toc25745406)

[2 Methodology 5](#_Toc25745407)

[3 Further Comments 11](#_Toc25745408)

# Introduction

OceanWise has a real-time environmental data management system named Port-Log. Port-Log is now used by many ports in the UK e.g. Peel Ports, ABP and the Port of London and increasingly by overseas ports and the energy sector globally.

The types of data handled by Port-Log include water level, water quality, directional and non-directional waves, wind and weather, current speed and direction, and air quality, although the system can be expanded to include other data types as necessary.

Port-Log receives data in real-time from sensors, manages it in a database and makes the data available via web pages and web services for use in maritime operations. Typical applications include Vessel Traffic Services (VTS), maritime pilotage, wind farm management and weather forecasting.

In addition to displaying and disseminating data in real-time, Port-Log stores data securely over time to provide a historical time series. Typical applications include engineering design and operational planning, survey correction, environmental assessment and sea level monitoring.

Port-Log customers have expressed an interest in making their data available for wider use outside their organisations. This action is being encouraged by governments to facilitate private and public sector data sharing and to enable coordination and more efficient use of resources.

Whilst private sector organisations have stated that they are willing to make their data more readily available, the process of doing so MUST be as simple and least time consuming as possible (e.g. pers comm Russel Bird, Hydrographer and Dredging Manager, Peel Ports).

The first step in facilitating the above is to make potential users know the data exists and once they know that it exists where and how the data can be accessed. It follows, therefore, that Port-Log should be able to create MEDIN compliant metadata that describes the Port-Log system and its contents, and to submit that metadata to the MEDIN repository.

In response to this opportunity, and to encourage private sector data sharing, MEDIN agreed to contribute funds on a 50/50 cost sharing basis to the design and development of a Port-Log discovery service that could be tested and implemented within MEDIN.

OceanWise will solicit permission from its Port-Log customers to enable the service, so the metadata can be made available. One such customer, Peel Ports, has already agreed to this in principle although the details of how this would work in practice still need to be agreed.

In addition, the results of this work will be published as a resource for organisations wishing to automate the generation of metadata from real-time data collection systems.

# Methodology

Port-Log is a live environmental monitoring data collection, management and display system to which new customers, stations and sensors are added on a regular basis. New customers are either added to an existing shared Port-Log server or provided with a dedicated server, which is cloned from a ‘source’ server and configured for their purpose.

The source server and hence its clones are updated with security patches and software upgrades on a routine basis. All servers are monitored continuously for performance, resources adjusted accordingly, and backed up.

In addition to new monitoring stations being added for new customers, existing Port-Log customers can add existing or new monitoring stations to their network. A station can comprise a single sensor, or a suite of sensors, reporting a variety of different data types and parameters. Stored vocabularies are used when and wherever possible, although there is potential for greater standardisation in the future.

Each sensor is associated with calibration record, which is a time bound i.e. the calibration is specified to be valid for a certain time/date period. During normal operations, calibration records are added to existing sensors e.g. if the sensor height changes, or for new, temporary or permanent replacement sensors. Calibrations records are never removed or amended but can be deprecated e.g. if found to be incorrect.

Details of the sensor calibration records are stored in each Port-Log database, along with sensor (instrument) manufacturer, model and serial number. Each sensor is associated with a station and each station is associated with a customer (who is deemed the data owner). An event log records all changes.

In order to be efficient, to align itself to the above approach and to meet the objectives of this work overall, the creation, maintenance and dissemination of metadata within Port-Log must be automated.

To achieve this aim, elements of each metadata instance must be stored in the Port-Log database and be capable of populating the metadata automatically, for example, in response to any of the changes described above. A new metadata instance is therefore generated when:

1. A new customer is added to Port-Log and the customer has agreed that the metadata can be disseminated (tick-box).
2. A new station is added by an existing customer
3. A new sensor (instrument) is added to an existing station
4. There is a change of sensor (replaced or removed)
5. A new calibration record is added to a new or existing sensor
6. Any other details are added or updated.

ISSUE 1 – Metadata that describe a dataset for discovery purposes will be created on a per site, per dataset, per calibration basis. A new calibration applied to a dataset will result in new dataset metadata that supersedes the dataset metadata for the old calibration. It does not overwrite it.

ISSUE 2 – So far, we have considered Port-Log metadata to be dataset metadata. Guidance is required on whether service metadata is relevant in this context. Metadata associated with time bound datasets downloaded from Port-Log would always be considered dataset metadata.

**Metadata Source Assessment**

In order to automatically create and update the metadata, it is necessary to have the required values stored in the Port-Log database.

Each metadata element contained in the MEDIN profile has been assessed against whether equivalent data is stored in the Port-Log database. The results of the assessment are presented in the table below:

|  |  |  |
| --- | --- | --- |
| Name | Notes | Is the content already in the database |
| Resource title | Each dataset in the database has a name which can be used to populate this | The resource title is formed from: <site name> - <dataset name> |
| Abstract | Description of the dataset | A means of recording an abstract in the database on a per site, per dataset, per calibration may be needed.At the moment it is created during metadata creation with the following template:Port-Log timeseries dataset measured at a site named <site> in <cluster>. The dataset name is <name>. The parameters available for the dataset are: <parameter list>. |
| Type | This will be 'dataset' for all records so it can be hard coded in the creation | The need for service metadata is bound to decisions on how the data will be provided to users. If it is provided by a service (e.g. web API) then that service could be described by discovery service metadata (see Issue 2 above). |
| Resource locator | If the dataset is available online then the URL for accessing the resource must be provided. This relates to the question of how online access can be provided to the data. | URL to be constructed on the fly during metadata creation (see Issue 2 above). |
| Unique resource identifier | This is an identifier of the dataset not the metadata. I think each dataset has a UUID which could be used in this case. This should be prefixed with a namespace which can be the internet domain of the owner/provider | Version 5 UUIDs are created from a namespace URI which is formed as:http://www.port-log.net/site/<site uuid>/dataset/<dataset name>This way they can be created in a repeatable way. But otherwise, the database does not have the concept of an identifier for a dataset that is stable and globally unique. |
| Resource language | This can be hardcoded in the metadata creation to English for the time being | Not needed – ENGLISH is entered. |
| Topic category | Each dataset must be categorized into one or more topic categories. The values are drawn from various BODC controlled vocabularies. | Tides, met, buoy, waves, water quality, CTD and currents = oceansMet and AQ = climatatologyMeteorologyAtmosphereRivers = inlandWater |
| Keywords | Each dataset will have one or more keywords. Keywords is a complex structure that includes the keyword, a URI since it'll be drawn from a controlled vocabulary and the citation of the vocabulary (title, publication date and type) | TBD base on parameters as for Topic Category |
| Geographic bounding box | This could be the extent covering the sites at which the dataset is collected | Site Location in Lat/Log (EPSG: 4326). Note single point monitoring stations only are being considered at this time. |
| Spatial reference system | At the moment this is WGS 84 only so it can be set at the metadata creation time | WGS84 (EPSG: 4326) – Port-Log is standardised on WGS84. |
| Temporal extent | The start date needs to be known. The end data may need to be known if data are no longer collected. If data collection is ongoing, this needs to be known | Dataset start and end date or null if ongoing. Also see note below. |
| Lineage | Text description of the sources used to create the dataset and how it came into being and so on. | See note below. |
| Access constraints | A code list value and text describing further meaning of the constraints (e.g. no constraints to access exist) | See note below. |
| Conditions for use | A code list value and text describing further meaning of the constraints (e.g. not to be used for navigation) | See note below. |
| Responsible party | There are a number of these, for which, at a minimum an email address and organisation name must be provided: originator, custodian, owner, metadata point of contact and distributor | Owner = CUSTOMERDistributor - OceanWise |
| Data format | Formats need to be covered at the moment are JSON and CSV. | Completed - hardcoded at the point of metadata creation - subject to testing by MEDIN |
| File identifier | This is the identifier of the metadata.  | Completed subject to testing by MEDIN |

The following additional issues are expanded upon from above table:

**Temporal Extent**

Mandatory element in the MEDIN metadata standard. The datasets that are being described are timeseries datasets. It is possible to encode the following information:

The start date - The first date that data landed in the database

The end date - The last date that data landed in the database (for datasets downloaded from PLG)

That data collection is ongoing or not - If data is not being actively collected then just the end date is encoded.

If data are being collected, then the end date is encoded in metadata and an attribute indeterminatePosition="after" is encoded to show that the end date is after that which is encoded.

**Lineage**

Mandatory element named lineage which describes the background information for a dataset. The definition is:

Lineage includes the background information, history of the sources of data used and can include data quality statements. The lineage element can include information about: source material; data collection methods used; data processing methods used; quality control processes. Please indicate any data collection standards used. Additional information source to record relevant references to the data e.g. reports, articles, website. Apart from describing the process history, the overall quality of the dataset or series should be included in the Lineage metadata element. This statement should contain any quality information required for interoperability and/or valuable for use and evaluation of the dataset or series.

A piece of text addresses this need created on a per site per dataset basis. Consideration must be given to whether a blanket lineage statement will suffice (i.e. one that will be applicable to all metadata instances) or whether lineage statements need to be individually tailored to each site/dataset combination.

The MEDIN definition included in the description provides help in how a statement might be framed, to include information about:

• source material (probably not relevant in the case of Port-Log)

• data collection methods

• data processing methods

• quality control processes

• data collection standards.

**Limitations on public access**

The definition is:

This element describes any restrictions imposed on the resource for security and other reasons using the controlled ISO vocabulary RestrictionCode. If restricted or otherRestrictions is chosen please provide information on any limitations to access of resource and the reasons for them. If there are no limitations on public access, this must be indicated i.e. No Restrictions on public access.

A piece of text that satisfies this definition is being included in the metadata. Bear in mind that metadata will be created on a per site per dataset basis so there will be many metadata records. Consideration needs to be given as to whether it is appropriate to have one blanket text or many, one per site per dataset basis.

This is encoded in the MD\_LegalConstraints::otherConstraints metadata element, if the MD\_LegalConstraints::accessConstraints is 'otherRestrictions'.

**Conditions applying for access and use**

The definition is:

This element provides information on any constraints on using the resource. Any known constraints such as licensing, fees, usage restrictions should be identified. If no conditions apply, then “no conditions apply” should be recorded.

If there is a formal licence title, that should be supplied along with, if available, a licence URL.

An appropriate piece of text which satisfies this definition is included in the metadata. Bear in mind that metadata will be created on a per site per dataset basis so there will be many metadata records. Consideration needs to be given to whether it is appropriate to have one blanket text or many, one per site per dataset basis.

The often-cited example is 'Not suitable for navigation'.

This is encoded in the MD\_LegalConstraints::useLimitation metadata element.

# Further Comments

At the time of issue of the current revision of this report, pilot metadata has successfully been produced and harvested from a Web Accessible Folder (WAF) on a Port-Log website. This metadata was successfully validated by the MEDIN core team.

In order to progress to the next stage i.e. look towards publishing metadata from Port-Log operationally then following need to be considered:

1. Receive and respond to feedback on the issues raised in Section 2 of this report.
2. Update the Port-Log software to allow the publishing of discovery metadata to be instigated when permission is received from data owners (tick box).
3. Review plans for updating UKDMOS – it is understood that this is done manually at present, whereas it should be possible to automate this as is the aim for Port-Log.
4. It is unlikely that data owners would wish to risk compromising their systems to over-use or cyber-attack (for example) as many are used in mission critical operations. Hence if data is to be published (and having come this far why wouldn’t it?) then a means to do so safely needs to be established.

This section to be completed following feedback from MEDIN on the pilot metadata and the issues described in this report.