Data and Services Discovery projects - Institutional Role in a Data Commons

Title

Integration of Northern Australia marine near real time meteorological and oceanographic data systems with the AODN

Approach

This project has made sensor data from remote weather/oceanographic stations operated by the Australian Institute of Marine Science (AIMS) more accessible via a cloud-based repository linking to the Australian Ocean Data Network (AODN) (https://portal.aodn.org.au/search?uuid=0887cb5b-b443-4e08-a169-038208109466). The entire archive of sensor data consisting of over 66 million observations are available and updated every few minutes when new data become available. The data are also accessible programmatically through an API and an R (a statistical data analysis) package. This allows data scientists using R to directly extract data for analysis in a repeatable and streamlined fashion. The API allows software engineers to dynamically extract data for use in visualisations and other applications. The AIMS weather web application (http://weather.aims.gov.au/) is an example of this.

The data are from 15 remote weather/oceanographic stations that transmit data every 10 minutes via a connected sensor network. Some of these stations have been operating for more than 30 years and provide useful historical time series for marine and climate research including coral bleaching, climatology and marine biodiversity studies.

Activities

The following activities were undertaken:

- Implement cloud database on Amazon RDS (a managed database service)Implement data synchronisation between cloud and on-premise databases
- Implement data synchronisation between cloud and on-premise databases. New data are collected every few minutes and these are synchronised to the cloud database after initial on-premise data QAQC processes are complete.
- Test synchronisation by uploading entire archive of over 66 million observations
- Enhance metadata to improve human readability, discoverability and to include AODN/ANDS vocabularies for instruments, parameters and organisations. Improve consistency of citation between data access mechanisms and across metadata record hierarchy. Utilise a DOI.
    - https://doi.org/10.25845/5c09bf93f315d
- Configure metadata and spatial services for integration with AODN Portal

    - https://portal.aodn.org.au/search?uuid=0887cb5b-b443-4e08-a169-038208109466
- Develop lightweight REST API and publish documentation
    - https://aims.github.io/data-platform/
- Develop R package wrapper for REST API to make it straightforward for R users to access the data programmatically
    - https://aims.github.io/data-platform-r/
- Develop R markdown notebook tutorial on using the R package
    - https://aims.github.io/data-platform-r/detailed-example.nb.html
- Migrate existing public facing AIMS weather station web application from Google Cloud to Amazon and refactor to use the REST API developed in this project.
    - http://weather.aims.gov.au/

Participants

The participants involved and their contributions were:

- AIMS Research Data Systems
    - Software and system design and implementation
    - Data management
- AODN
    - Feedback on dataset configuration for integration into AODN portal
    - Feedback regarding refinement of metadata information including title, credits, citation format and parameter vocabulary
- ARDC
    - DOI and citation clarification, particularly in context of parent and child metadata record structures
    - Progress review of work done
- AIMS R Users Group
    - Feedback on usability of the R package and associated R markdown tutorial

Outputs

A software system and associated artefacts were developed that provides enhanced access to this dataset in several fit for purpose ways:

- Researchers are able to filter, subset and download data targeted to their needs via the AODN portal user interface.
- Data scientists can use an R package for integrated, interoperable and repeatable access to the dataset directly in the R data analysis environments. An example R markdown notebook demonstrates the use of the R package. It shows how to access the data and shows several visualisations of the data.

- Software engineers can access the data via a lightweight REST API to build new applications, tools, visualisations etc in addition to the spatial web services provided by the AODN portal (WMS and WFS)
- The general public can access the latest observations via the web application which is underpinned by the REST API
- Enhanced metadata improves the discoverability of the dataset via web search and within various data catalogues including AIMS, AODN and RDA. The dataset is now citable using a Digital Object Identifier (DOI) minted using the ARDC service.

FAIR

A post project FAIR assessment has been completed and available in appendix A. The following is a summary of the improvements:

- Improved metadata, enhancing searchability and discoverability
- Use of AODN/ANDS vocabulary for standardisation on keywords
- API and R library for data access, analysis repeatability and integration
- AODN portal access and the use of standards for spatial data interoperability
- Use of a DOI and making it easier for data users to cite the dataset. Citation information is now easier to locate in each of the data access mechanisms.

Our ARDC Project Coordinator provided the following commentary:

“The End of Project information is explicit and shows where the data is more FAIR as a result of the project. Every component shows elements of improvement, except Reusable, where you were already achieving gold standard for licensing.
Provenance capture may be worth in the future if, as you say, resources allow. We conducted a survey on data trustworthiness recently and almost all responses said data provenance was a key factor for reuse.
In the main, the project is a great success and this assessment attests to that.”

Collaboration

- Aspects of the project that have been jointly developed with the ARDC include creation of DOI and citation statements within a set of structured, related metadata records
- We expect researchers, data scientists to access the data via the mechanisms provided by the project, however, we are unable to report on this at this time as it has only just been operationalised.
- Improved collaboration relationship between AIMS and AODN and ARDC
- Contribution of the data to the AODN Portal was a collaborative output between AIMS and the AODN. The focus was on metadata conformance and content improvement along with conformance to spatial data standards.

Sustainability

The following ingredients ensure future sustainability:

- The systems in place perform business as usual functions at AIMS and as such are self-sustaining
- AIMS has an internal software engineering/data management team responsible for this system
- AIMS policy for open data access ensure the data will continue to be made available for reuse

The data collection is sustainable for the following reasons:

- Future projects will utilise the cloud infrastructure and database put in place by this project, placing it at the core of AIMS' cloud-based data management systems
- This data is an ongoing collection of significant importance (unique time-series for climate study's and as ancillary data for marine research activities). It forms a key part of AIMS monitoring activities and will continue into the foreseeable future.

Learnings

There were challenges in handling this dataset's complexity with differing ownerships and funding bodies for various parts of the dataset which impacted how it should be cited and attributed and introduced complex metadata hierarchy to reflect the aggregate nature of the dataset. At this stage, a balanced approach was used to publish the dataset under a single DOI with a consistent citation format and a consolidation of information across the metadata hierarchy.

There is opportunity for the ARDC to play a part in the integration of institutional data infrastructure with national infrastructure by:

- Assisting in efforts that capture dataset provenance information, particularly as part of data quality control and assurance processes.
- Continuing to assist in efforts to make institutional data sets FAIR with focus on availability via API/web services.
- Assisting in efforts to establish or promote data standards for these web services, including time series data.
- Assisting institutions in the development and uptake of national web services for standardised data processing, particularly for the quality control and assurance of data from commonly used instrumentation.

Impact

The project has had the following impact on research efficiency:

- Data are easier to find and extract and use, particularly when using the R package.
    - When using the R package, the researcher will be able to access the latest data with minimal effort when compared to the traditional find and download approach.
- AODN portal allows extraction across sites very easily
- Researchers can more easily cite the data they are using

The project has improved research integrity in the following ways:

- The R package allows for repeatable data analysis. The researcher can re-run their analysis.
- DOI's allow for citation of the dataset and provide a persistent identifier.

The data contain commonly sought environmental parameters over extended time series across a large remote spatial area in near real time and as such have a large base of diverse end-users including researchers, software developers, data analysts as well as tourism, recreational, law enforcement and industry. The choice of data access via AODN Portal, public facing website and programmatic API cater for the specific needs of each of these user groups.

Report prepared by: Mark Rehbein, Australian Institute of Marine Science
Date: 24/09/2019