ROIS Final Report

Executive Summary

The Polar Environment Data Science Center (PEDSC) will become increasingly relevant as a base source of information for evidence-based decision making related to the Antarctic within Japan, but also for the international community.

Areas of potential mutual engagement with Australia include:

- **Virtual Database and AI**;
- **Digital Earth Antarctica**;
- **Data to Support Operations and Logistics**;
- **Seabed Mapping and Marine Data**;
- **Space Interests**; and
- **Participating in a Staff Exchange Program with Australia**

Potential challenges for the PEDSC include:

- **Lack of Dedicated Data Systems Support Staff**;
- **Limited Co-ordination of Data Management Initiatives Across ROIS-DS**;
- **Increasing Data Scale and Complexity and Requirements for Data Standards**; and
- **Management of Data Collected by Japan's New Icebreaker**

Recommendations include:

- **Increasing Recruitment of Dedicated Data Staff**;
- **Clarifying Lines of Responsibility Relating to Data**;
- **Increasing Program (whole of ROIS)-level Communication and Coordination**;
- **Increasing Japan's Leadership in International Fora**;
- **Gaining World Data System CoreTrustSeal Certification**; and
- **Increasing Digitisation of Physical Data Collections**

Potential **institutions for comparison** include:

- **The Australian Antarctic Data Centre (AADC)**;
- **The British Antarctic Survey (BAS)**;
- **PANGAEA**; and
- **The Ocean Biodiversity Information System (OBIS)**

Introduction

As part of their international advisor program, the Research Organization of Information and Systems (ROIS) invited Dr. Johnathan Kool of the Australian Antarctic Division's Data Centre to engage with staff at the Polar Environment Data Science Center (PEDSC) at the National Institute of Polar Research (NIPR) in Tachikawa, Tokyo. The visit was partly motivated as a result of existing engagement through the Standing Committee on Antarctic Data Management (SCADM). The aims of the advisory role were to develop an improved mutual understanding of data management systems and practices, with an aim towards improved collaboration on projects and programs.

Current State

PEDSC Administrative Structure

The PEDSC is managed as part of ROIS's Joint Support Center for Data Science Research (DS), although it is physically located on the NIPR's Tachikawa campus. The PEDSC is managed by Akira Kadokura (family names are ordered last in this document), and currently has a total of 6 principal researchers, and an additional 9 concurrent researchers. Three of the staff (Akira Kadokura, Yoshimasa Tanaka, Masayoshi Kozai) are associated with atmospheric physics/astronomy, two with geophysics/earth science (Masaki Kanao, Jun'ichi Okuno), and one with bioscience (Kunio Takahashi). A published overview of the PEDSC is available at http://doi.org/10.5334/dsj-2022-012.

The larger DS organisation is led by Professor Hiroyuki Araki. Other centers within the DS include the Database Center for Life Science, the Center for Social Data Structuring, the Center for Open Data in the Humanities, the Center for Genome Informatics, and the Center for Data Assimilation Research and Applications. There is also a Data Science Promotion Section, which is responsible for media and outreach.

The NIPR is led by Director-General Takuji Nakamura, with Dr. Yoshifumi Nogi serving as the Vice Director-General. The NIPR is co-located with the Institute of Statistical Mathematics, and the National Institute for Japanese Language and Linguistics (in a separate building).

PEDSC Data Storage and Infrastructure

The PEDSC maintains its data collections in different databases managed by individual programs and researchers. For example, the terrestrial biodiversity database is maintained independently of seismic data, or radar information. Oceanographic data is maintained separately from the NIPR by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) and the Japan Meteorological Agency (JMA). NIPR Data is stored on servers located on the lower floor of the NIPR building, however there are additional servers in other areas of the building. The NIPR's computer and networking infrastructure are maintained by the Information and Communication Center, managed by Director Masaki Okada. Like the Australian Antarctic Division's ICT Section, the Information and Communication Center is responsible for communications with stations. Video conferencing to Syowa Station is common, and typically bandwidth out from Tachikawa is approximately 7 Mbps, and station bandwidth is approximately 4 Mbps. Quality of Service (QoS) is organised by priority group.

Physical collections

The NIPR retains a number of physical collections relating to data and physical samples, including collections of meteorites collected in Antarctica, heritage items, as well as paper sheets and magnetic tapes containing data records. Analogous to the Australian Antarctic Division's relationship with the National Archives of Australia, the NIPR sometimes transfers custody of items having high and enduring value to the National Archives of Japan. There is also a permanent Polar Science Museum that displays Antarctic heritage items in a dedicated building located on the Tachikawa campus.

GIS and Remote Sensing

The NIPR's GIS information is principally delivered through its ADS web interface (https://ads.nipr.ac.jp/antarctic-gis/). The web page delivers spatial information as download links for individual data collections associated with metadata entries. OGC Web services for this information have not been developed at this time. Remotely-sensed data is chiefly made available through JAXA (e.g. G-Portal: https://gportal.jaxa.jp/gpr/?lang=en). Management of remotely-sensed data is based on project, and has not been centralised.

Topics of Joint Interest and Potential Collaboration (Australia-Japan)

Virtual Database and AI

The AADC is currently engaged in a program through the Centre for Antarctic and Southern Ocean Technology (CAST) to develop 'Virtual Database' capability. The 'Virtual Database' leverages machine-learning/artificial intelligence as a tool to assist with understanding the structure and nature of data sets. It does this by reading both metadata and data, and then attempting to interpret data content, for example, by interpreting field header information. Data content is then linked to controlled vocabularies (e.g., the NERC vocabulary server) so that data contents can be read and understood through code. Having machine-readable data structure helps to provide better data discovery and interoperability. Variables can be grouped and faceted for searching and browsing, as well as referencing linked concepts. Similar variables (e.g., temperature) can also be transformed as required, or excluded if there is an important distinction between them. Using AI helps provide completeness and consistency for metadata, and eliminating a highly tedious and time-consuming task for data managers.

The AI-system can be further developed to assist researchers in completing metadata associated with data sets, prompting them for additional keywords, and potentially pre-filling sections as required. If the system receives feedback that its suggestions are incorrect, it will update its processes to reflect the new information.

The AADC intends to develop this capability as an open-source project. The NIPR would be free to incorporate the application into its own workflows if desired.

Digital Earth Antarctica

The AADC will be undertaking a project, in conjunction with Geoscience Australia, to develop and deliver analysis-ready earth observation data for Antarctica. The work will leverage existing capability developed through the Digital Earth Australia and Digital Earth Africa programs. As a platform, Digital Earth can be used to generate a broad range of analytical products, including time-series analysis, regional statistics, and trend detection. The technology was initially developed to use surface reflectance data, but additional forms of data can potentially be incorporated as well (e.g., radar and lidar). As with the Virtual Database, Digital Earth technologies are open source. NIPR could either use the platform developed by Australia to develop products, or it could develop its own platform and processes.

Data to Support Operations and Logistics

Both Australia and Japan have a strong interest in undertaking safe, efficient, and effective operations. Both maintain station infrastructure and polar vessels, as well as undertaking significant cargo operations, and both programs would benefit from overall improved situational awareness. Data interoperability will be important for ensuring the greatest possible coverage and use of information by international partners, and may also support new operational support capabilities (e.g. increased telepresence).

During discussions with Dr. Gen Hashida, the importance of continued delivery of the COMNAP Asset Tracker System (CATS) was discussed. CATS currently receives development support through the AADC, however, presently there are no staff dedicated towards this work. Instead, work is performed on an 'as needed' basis. The AAD have already committed to recruiting a permanent, ongoing Operational Mapping Support Officer, and it would be reasonable to expect that the responsibility for administering CATS would be assigned to the individual taking on that role. Dr. Hashida agreed to raise the importance of maintaining the CATS system at an upcoming COMNAP meeting to promote

the work. There is also an opportunity for expanding the scope of CATS to include a broader range of assets (e.g. type and source country). Existing capability of the system allows for including assets positioned anywhere in the region; the system only requires accurate positional information feeds. Positional information from Japan would be welcomed.

There is also mutual interest in increased communications bandwidth to support data transfer from the continent, but also on the continent (e.g. inter-station transfers, wireless network support for field programs). There may be opportunities for developing shared data networking capabilities. For example, ships and stations could serve as nodes within a collective wireless information network, allowing for greater shared data transfer capability from remote field devices.

Seabed Mapping and Marine Data

Seabed mapping and marine data are primarily associated with JAMSTEC. However, the NIPR and PEDSC will need to consider how data collected from JAMSTEC's icebreakers will be curated and delivered, particularly given that Japan is set to be bringing a new icebreaker into service within the span of a few years.

It will also be important to consider how Japan's Antarctic multibeam sonar data can be integrated into global collections in an ongoing manner. The Australian government has developed a program 'AusSeabed' with the aims of collating, processing, and integrating Australia's collection of seabed information. As part of this, the project has been developing open-source technologies, which could potentially be used by Japan, and combined with Australian information. In turn, this would contribute to initiatives such as the Nippon Foundation-GEBCO Seabed 2030 project.

Space Interests

Japan is extensively engaged with space research through a range of scientific activities, as well as collaboration with the Japan Aerospace Exploration Agency (JAXA). In addition to the aforementioned research on meteorites (including samples retrieved by Hayabusa 2), Japanese researchers are engaged with ionospheric measurements, drone research for retrieving lunar samples, and space weather measurements, among others. There is a strong and clear intersection between space research and the activities of the NIPR, and there is considerable scope for the AAD to learn from the strengths of these programs.

From JAXA, there was strong interest in joint participation if Australia were to develop earth observation ground station capability at one of its stations. There are also potential opportunities for joint field observation programs to support calibration and validation of satellite observations. The Group on Earth Observations (GEO) may provide a forum for focused development of joint requirements, or alternatively the SCAR Open Science Conference in 2024, and the AAD's Remote Sensing Data Manager has now joined the Group on Earth Observation's Cold Regions Initiative group (GEOCRI).

Potential Challenges for ROIS-DS and the PEDSC

Lack of dedicated data systems support staff

PEDSC staff are both qualified and highly capable, however most are associated with a science research background as opposed to having a dedicated IT or data management background. There are significant challenges relating to data system design and management that require training and experience in order to take full advantage of modern best-practice solutions. Currently, PEDSC staff must balance progressing research goals with the development of data systems and associated management responsibility. The PEDSC would benefit considerably from having staff whose sole

function would lie in developing and maintaining data systems. Rather than focusing on a research domain, corporate staff should exist whose purpose it would be to think about system-level design and delivery. Dedicated software developers and engineers would be required to implement solutions, and in turn, some form of project management capability may be required to ensure that tasks are proceeding according to schedule.

Limited co-ordination of data management initiatives across ROIS-DS

Although there is active communication between different projects and programs within the NIPR and ROIS, there are opportunities for improved coordination. For example, IUGONet, AMIDER (part of PEDSC), and the ADS (managed within NIPR) share similar objectives relating to improved data delivery, but appear to have been progressing independently. Each project has its own individual value, but coordinating their development would likely deliver even greater value. Similarly, there could be strong value in having greater mutual visibility among programs within ROIS. For example, solutions developed for social data by the Social Data Structuring Center, for genetic data by the Genome Data Analysis Support Center, and especially biological data by the Life Science Integrated Database Center, could all be of potential relevance to the PEDSC. Similarly, the nearby National Institute for Japanese Language and Linguistics might have some tools and techniques relating to natural language processing and data mining from documents. The PEDSC and all of ROIS stands to gain considerably in terms of productivity and efficiency from increased internal communications and engagement.

Increasing Data Scale and Complexity and Requirements for Data Standards

As PEDSC's data collections continue to expand in terms of size and complexity, it will become increasingly important to develop ways of ensuring that the data are interoperable and consistent. Currently, the PEDSC's information is maintained in separate databases associated with individual projects and research programs. Distributed data management can be an effective approach, but it will require increased attention to the use of standard naming (controlled vocabularies), and well-understood (ideally, machine readable) data structures and formats. Interoperability requirements would also potentially extend to data feeds from Japanese sensors located at Australian stations.

Management of Data Collected by Japan's New Icebreaker

As part of commissioning it's new icebreaker RSV *Nuyina*, the AAD identified a need to develop strong processes relating to the collection and curation of data from this state-of-the-art asset. *Nuyina* represented a step-change in terms of capability, with commensurate increases in the volume and diversity of information relative to its predecessor, RSV *Aurora Australis*. Currently, Japan is in the process of building its own new icebreaker, and will likely be facing similar challenges. Although the vessel is being managed by JAMSTEC, the PEDSC should consider how it will be accessing this data, and presenting requirements to ensure interoperability with its systems and other data collections.

Participating in a Staff Exchange Program with Australia

There are a number of opportunities for technical exchange, particularly relating to engineering and robotics. Japan's development of super-pressure balloons as a platform for sensor monitoring would be an area of potential collaboration, as would Improvements in GNSS availability and positioning capability (as developed by Jun'ichi Okuno). Additional exchange opportunities could include shared mapping opportunities under Australia's East Antarctic Mapping Program, or staff exchanges aboard icebreakers.

Recommendations

Increase Recruitment of Dedicated Data Staff

ROIS should consider resourcing additional staff at the PEDSC with the sole responsibility of developing and managing data systems. If some sort of research focus is essential, then the focus of this research should be data science. Their primary focus should be thinking at a 'system' level, rather than specialising in a particular scientific domain or application. To support this work, dedicated staff would also be needed to develop data infrastructure (i.e. data 'engineers'), data interfaces (web services and delivery systems), and undertake general software coding and data management tasks. At a minimum, one high-level dedicated data manager should be recruited, along with a team of 2-3 data engineers/software developers. Ideally, this number could be increased to include data records managers (responsible for low-level file management), and additional data infrastructure staff as required to scale appropriately.

Clarify of lines of responsibility relating to data

As the responsibilities of the PEDSC increase, it will become increasingly important to develop clear lines of responsibility for different individuals and groups within the PEDSC, NIPR, and ROIS-DS. It may be worthwhile developing a concrete Data Policy document that includes roles and responsibilities – which is what has been done in the AADC. Similarly, it may be useful to develop a Data Strategy document that details the current technology and data delivery landscape within the PEDSC, as well as a roadmap for moving forward into the future. As part of developing these documents, a diagram of responsibilities could be developed highlighting the different components of ROIS and external partners (e.g. JAMSTEC, JAXA) to provide a better understanding of interdependencies, and to assist in streamlining data delivery systems.

Increased Program-level Communication and Coordination

The PEDSC should consider mechanisms for increasing engagement with other data groups within ROIS, as well as moving towards more strategic, program-level interaction with other government agencies such as JAMSTEC and JAXA. Currently, engagement appears to take place at the level of project, but a broader program-scale (i.e. PEDSC – Life Science Integrated Database Center) coordination should also be considered. Mechanisms for increasing data connectedness and interoperability should be considered, including identifying common standards such as controlled vocabulary terms and open data formats.

Also relating to communication and coordination, the PEDSC should begin engaging with JAMSTEC to identify roles and responsibilities relating to data pipelines for Japan's new icebreaker. Even if all data collection and management aboard the vessel will be handled solely by JAMSTEC, a mechanism should still be developed for ensuring that data are findable and accessible through the PEDSC.

Increase Japan's Leadership in International Fora

Although Japan and the PEDSC have strong engagement with the international community, there are opportunities for increased leadership opportunities on the international stage. Japan has a strong reputation internationally, particularly through its prior hosting of the World Data System program, and is well-placed as a key country in the Asia-Pacific Region. Individuals who are able to take on leadership roles at the international level should be developed and supported, and this may also require allowing for dedicated time to fulfil these responsibilities rather than being on top of existing workloads. Potential examples could include assuming leadership roles within the Standing Committee for Geographic Information (SCAGI) under SCAR or the Southern Ocean Observing System (SOOS), among others.

Gain World Data System CoreTrustSeal Certification

Especially in light of Japan's long-standing connection with the World Data System (WDS), the PEDSC should follow through with its plans to seek formal CoreTrustSeal Certification. The process of going through the accreditation will help the PEDSC identify concrete steps to conform with current international best practices, and will serve as independent confirmation of the stability and security of its systems.

Increase Digitisation of physical data collections

Although the NIPR has abundant physical space for storing collections, over time, the magnetic tape will be at risk of physical degradation, in addition to decreasing availability of hardware capable of reading the media. ROIS should consider increasing support for digitising its physical archive. Not only would this have the benefit of preventing potential data loss, but it would also increase the accessibility and interoperability of the archived information. Furthermore, less space would be required to maintain the collection, and it would be consolidated with other digital assets.

Potential institutions for comparison

With regards to data management, different institutions will have different needs based on their mission, size, complexity, and growth trajectory. For example, the AADC does not have the same research focus that the PEDSC does, nor does it have the requirement to have an Arctic focus as well as Antarctic.

Although a number of organisations can be presented as examples of good data management practice for the PEDSC, it is important to recognise that there may be very good reasons for the PEDSC to adopt and maintain its own particular style of data management according to its needs. Bearing this in mind, a list of some potential benchmark organizations is provided below.

The Australian Antarctic Data Centre (AADC)

The AADC currently has over 25 staff, having recently increased its number following a sustainable funding review. The AADC is composed of 3 principal sections – Mapping & GIS, Infrastructure & Applications, and Information Management. The Mapping & GIS section is responsible for not only data and map delivery, but also coordinating all aspects of mapping and surveying programs in Antarctica. Infrastructure & Applications are responsible for software and systems engineering to support online delivery of data. Information Management is responsible for the AAD library, records management, and handling of data from RSV *Nuyina*.

The British Antarctic Survey (BAS)

BAS have a strong reputation for data delivery, including the successful delivery of their new BEDMAP products. At Halley Research Station, BAS have installed a number of automated scientific containers that can be remotely operated. Emulating this capability has been an area of interest for the AADC.

PANGAEA

Pangaea also has a very strong reputation for data delivery as well as ease of use. Their website is well-organised, and contains a large, diverse, and comprehensive collection of data sets.

Ocean Biodiversity Information System (OBIS)

Although OBIS only handles marine biological information, it handles a broad variety of information at a global scale. OBIS is also tightly integrated with the World Register of Marine Species (WoRMS), which provides well-controlled taxonomic names.

Other potential institutions to note are the Istituto di Scienze Polari of Italy, who have been making significant strides to improve their data delivery systems, and similarly, the Swiss Polar Institute.

Dr. Johnathan Kool
Australian Antarctic Data Centre Manager
Australian Antarctic Division, Department of Climate Change, Energy, Environment and Water
203 Channel Highway, Kingston, TAS 7050