Activity Report from International Strategic Advisor

POLAR ENVIRONMENTAL DATA SCIENCE CENTRE (PEDSC)

20th November 2023 – 18th December 2023

Prepared For: Joint Support – Centre for Data Science Research (ROIS-DS)

Prepared By: KASSIM S. MWITONDI (PhD)

January 2024

Table of Contents

List of Figures

Executive Summary

This report is prepared for the President of Research Organization of Information and Systems (ROIS) and the Director of the Joint Support-Centre for Data Science Research (ROIS-DS). It outlines a number of advisory activities that the author accomplished as an invited visitor to ROIS-DS' Polar Environment Data Science Centre (PEDSC) at the Tachikawa campus between 20th Nov and 10th Dec 2023. It provides inputs to the management, administration and future direction of PEDSC/ROIS-DS focusing on the current evaluation and how it potentially leads to future research directions within the research network. Based at the PEDSC, the author had interactive sessions and gained familiarity with research activities at the following centres.

1) Polar Environment Data Science Centre (PEDSC)
2) Centre for Data Assimilation Research and Applications (CDARA)
3) Centre for Open Data in the Humanities (CODH)
4) Database Centre for Life Science (DBCLS)
5) Centre for Social Data Structuring (CSDS)

The main objective of those sessions was to gain insights into the research activities at each centre, assess inherent potential within each, evaluate the current status and make recommendations for future directions on ROIS-DS' enhancement of data science applications across the network. This objective leads to the main recommendation made by this report - namely, ***aligning with global trends of data-intensive research***. The recommendation is based on a relevant SWOT analysis (Stewart & Benepe, 1965) from an international point of view.

It highlights key steps to be taken to consolidate existing international collaborations that is already taking place across within the network and identifies PEDSC, CDARA, CODH, DBCLS and CSDS as non-orthogonal drivers of data-driven research transformation. The report also identifies a number of benchmarking institutions for PEDSC's unifying role in the ROIS-DS network, based on which it makes the following specific recommendations.

1) Industrial placement schemes for higher education students: This will help address time-consuming issues relating to data-deluge challenges such as data harmonisation. The scheme is also a great way for imparting knowledge on young researchers and it helps alleviate the impact of data science skills that the centres face.
2) Cloud-based data sharing: Inspired by the AMIDER database, currently being developed at the PEDSC, this initiative provides a potential for sharing structured and unstructured data in an integrated cloud and Internet of Things (IoT) environment.
3) Tapping into the CDARA potential: It was noted CDARA's modelling potential wasn't fully utilised across ROIS-DS and that in combination with recommendations 1 and 2 the network can achieve a well-organised environment for sharing tools, data and skills.
4) ROIS-DS research centres to draw lessons from benchmarking institutions that conduct related research and adapt novel technologies from data-intensive practitioners.

The report confines itself to the foregoing recommendations, providing generic comments about ROIS-DS' management and future plans. That means, the report may be used, in the future, as input to any relevant work with full access to ROIS-DS' strategic and/or budget plans.

Section 1 Introduction

The report outlines a number of advisory activities that the author accomplished as an invited visitor to ROIS-DS' Polar Environment Data Science Centre (PEDSC) at the Tachikawa campus (20th Nov-10th Dec 2023). The visit was planned to coincide with the *International Symposium on Data Science 2023 (DSWS-2023): Building an Open-Data Collaborative Network in the Asia-Oceania Area*, that took place in Tokyo from 11th to 15th December 2023 on which the author was a member of its International Advisory Committee (IAC). The symposium was part of PEDSC's – hence ROIS-DS' continuum of data science related research activities – a key motivation for its establishment (ROIS-DS, 2016).

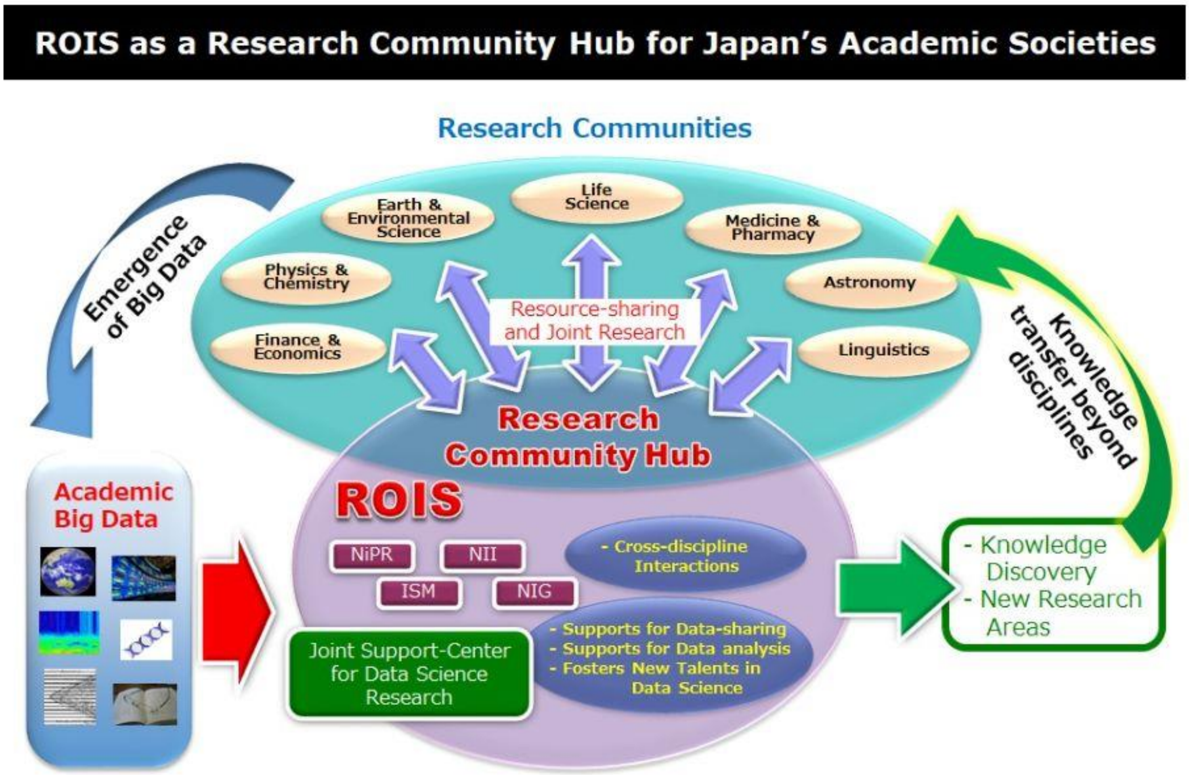

ROIS as a Research Community Hub for Japan's Academic Societies
Research Communities
Earth & Environmental Science
Life Science
Medicine & Pharmacy
Physics & Chemistry
Astronomy
Finance & Economics
Linguistics
Resource-sharing and Joint Research
Emergence of Big Data
Knowledge transfer beyond disciplines
Research Community Hub
ROIS
NiPR
NII
ISM
NIG
- Cross-discipline Interactions
Academic Big Data
Joint Support-Center for Data Science Research
- Supports for Data-sharing
- Supports for Data analysis
- Fosters New Talents in Data Science
- Knowledge Discovery
- New Research Areas

Figure 1: ROIS' structure for data-driven and knowledge generation initiatives

Figure 1 graphically illustrates ROIS' as a research hub for different academic societies in Japan, in which it can be seen that the following five national institutes fall under ROIS's remit:

1) National Institute of Polar Research (NIPR)
2) National Institute of Informatics (NII)
3) Institute of Statistical Mathematics (ISM)
4) National Institute of Genetics (NIG)
5) Joint Support-Centre for Data Science Research (ROIS-DS)

ROIS-DS's main objective is to support academic and industrial research, including that of students, to conduct data-driven scientific research. The report is prepared for the President of Research Organization of Information and Systems (ROIS) and the Director of the Joint Support-Centre for Data Science Research (ROIS-DS). It outlines a number of advisory activities that the author accomplished as an invited visitor to ROIS-DS' Polar Environment Data Science Centre (PEDSC) at the Tachikawa campus between 20th Nov and 10th Dec 2023.

It is designed to provide inputs to the management, administration and future direction of the network, focusing on the current evaluation and how it potentially leads to the network's future research directions. It particularly emphasises research management, education and international collaboration with external research communities, and it makes logistical and strategic recommendations. During the time spent at the Tachikawa campus, the author had interactive sessions with the following research centres, associated with the institutes above.

1) Polar Environment Data Science Centre (PEDSC)
2) Centre for Data Assimilation Research and Applications (CDARA)
3) Centre for Open Data in the Humanities (CODH)
4) Database Centre for Life Science (DBCLS)
5) Centre for Social Data Structuring (CSDS)

The main objective of those sessions was to gain understanding of the research activities at each centre, assess inherent potential within each, evaluate the current status and make recommendations for ROIS-DS' future directions on data science applications across the network. This objective leads to the main recommendation made by this report - namely, ***aligning with global trends of data-intensive research***. The recommendation is based on a relevant SWOT analysis (Stewart & Benepe, 1965) from an international point of view.

It highlights key steps to be taken to consolidate existing international collaborations that is already taking place across within the network and identifies PEDSC, CDARA, CODH, DBCLS and CSDS as non-orthogonal drivers of data-driven research transformation. The report also identifies a number of benchmarking institutions for PEDSC's unifying role in the ROIS-DS network, based on which it makes the following specific recommendations.

1) Industrial placement schemes for higher education students: This will help address time-consuming issues relating to data-deluge challenges such as data harmonisation. The scheme is also a great way for imparting knowledge on young researchers and it helps alleviate the impact of data science skills that the centres face.
2) Cloud-based data sharing: Inspired by the AMIDER database, currently being developed at the PEDSC, this initiative provides a potential for sharing structured and unstructured data in an integrated cloud and Internet of Things (IoT) environment.
3) Tapping into the CDARA potential: It was noted CDARA's modelling potential wasn't fully utilised across ROIS-DS and that in combination with recommendations 1 and 2 the network can achieve a well-organised environment for sharing tools, data and skills.
4) ROIS-DS research centres to draw lessons from benchmarking institutions that conduct related research and adapt novel technologies from data-intensive practitioners.

The next exposition provides an outline of the interactions the visitor had with each of the foregoing research centres. It is worth noting that research activities at each of the centres are far broader than those addressed in this report, which are confined to specific research activities that were either discussed, presented or demonstrated during the visits.

Section 1.1: Polar Environment Data Science Centre (PEDSC)

The Polar Environment Data Science Centre (PEDSC) promotes access and sharing of the scientific data that is routinely collected via scientific observations and research activities in the Antarctic and arctic regions by the polar science research community. The PEDSC was formally launched in 2017 and it has since worked in close collaboration with universities and other external communities, based on the standard academic ranking staffing structure. As part of ROIS-DS, the centre supports archiving, processing, analysing and sharing of polar environment data collected by scientific missions and observatories in the Antarctic and Arctic regions. The following are the visitor's comments on PEDSC and related centres.

Section 1.1.1: Data deluge at PEDSC

The author visited the space and upper atmospheric sciences, the meteorology and glaciology and the geosciences centres. While those three constituted only part of a much larger data storage capacity, it was evident that the PEDSC generates and has access to an awful lot of data, far beyond the centre's capacity to process it. For instance, the upper atmospheric sciences centre receives live data transmission from NIPR's monitoring centre, in the Antarctica and at the PEDSC head office there is live streaming of seismological data. Apparently, PEDSC is not sufficiently staffed to optimally manage and utilise the data inflow. The author understands that such data is readily available to scientists and researchers across Japan and to all collaborating cold region partners. There is potential for that practice to be extended to include data assimilation applications run from within ROIS-DS, using for instance, postgraduate students affiliated with the Institute of Statistical Mathematics (ISM).

Other examples include the meteorite storage room that has over 20,000 stored meteorites, collected from the Antarctica and the ice cores at the low temperature room where samples collected mainly from the Antarctica are stored. Both the meteorites and the ice cores contain highly useful chemical composition data, with potentially highly useful information, much of which remains buried in the samples. For instance, ice cores contain air bubbles some dating back to more than 10,000 years, which may be quite informative about current surface temperature levels (Buizert, 2021). Classification of meteorites presents another challenge and opportunity as they are currently identified and classified after complex and slow chemical analyses processes. The centre currently stores them into three categories: achondrites, ordinary chondrites, and carbonaceous chondrites. One way of enhancing performance at the meteorites and the ice cores centres would be to automate analytical processes by adopting Machine Learning (ML) techniques, for instance. In discussions with staff at both centres, the need for intensive data analysis skills was evident but so were staffing constraints.

Section 1.1.2: Staffing Constraints

Staff at both centres were fully aware of the potentially useful data that remains buried in the samples but, much as they would have liked to get out and share data assimilation results, they were seriously constrained on staffing. Different options of addressing this issue were discussed – including deploying Machine Learning (ML) models for automatic detection of features in samples. It was noted that some of these ideas had previously been explored. At the meteorites centre, for instance, the author was advised that a third member of staff, who was away on a scientific mission in Antarctica during that visit, had been very keen in deploying such techniques. Both centres expressed great eagerness to pursue the idea further. Recommendations on how the institute can proceed in this respect are made in Section 2.

Section 1.2: Centre for Data Assimilation Research & Applications (CDARA)

The visit to CDARA highlighted both challenges and opportunities as one would normally encounter in the data modelling world – that of merging computer generated information and real-life observations. CDARA adopts the historical application patterns by focusing on atmospheric and weather research (Zhang et. al, 2020), as evidenced by the presentation at the centre and at the Centre for Open Data in the Humanities (CODH). In both presentations, the underlying problem was detection of natural structures in data, for which finite mixtures models (McLachlan and Peel, 2000) were applied. This section highlights the key points that emerged from the discussions at CDARA, the CODH application is discussed in Section 1.3:

Apparently, CDARA's major strength lies in its modelling skills capacity – particularly the application of statistical and data science tools and techniques that can readily be adapted to other centres. While there is a well-established working relationship between CDARA and CODH and between CDARA and the National Institute of Genetics (NIG), it looks like such collaboration has not been extended to other centres. Given the amount of data that PEDSC generates, a sustained relationship with CDARA would lead to profound scientific research outcomes. The same can be said about links to DBCLS and CSDS. Section 2 makes recommendations on how CDARA can significantly contribute to other centres.

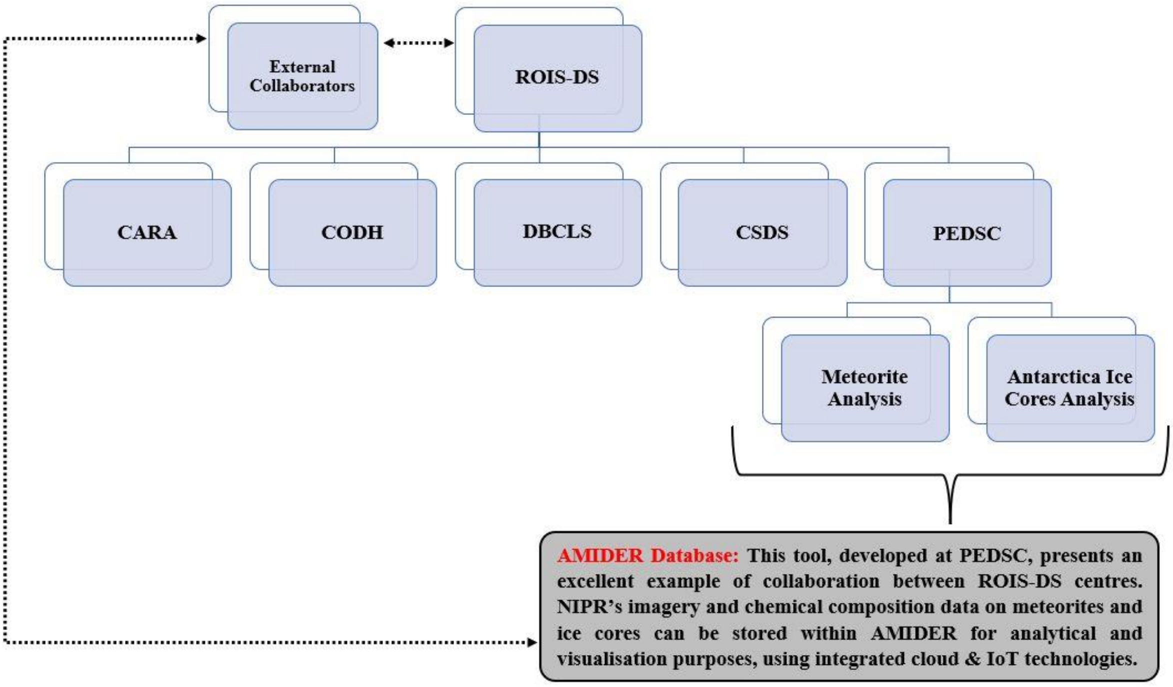
Section 1.3: Centre for Open Data in the Humanities (CODH)

The CODH seeks to promote data-driven research and external collaboration in various areas of humanity research. It is embarking upon developing science-driven, digitised humanities research across organizations within Japan and beyond. The centre promotes open access to data providing access to humanities data using state-of-the-art technologies and modelling techniques. It was quite intriguing to see the large volume of real-life applications CODH has its eyes on, which is great in promoting interdisciplinary research. However, that also implies that the centre requires a diverse range of skills and staffing not only of data professionals, but also of relevant underlying domain knowledge areas. CODH's current direction is in deepening Machine Learning (ML) applications with an increased Citizen Science (CS) content, and they have been looking for partners both in terms of technology and problems space (in humanities).

The CODH staff showcased some of their on-going research work and three research application tools stood out – historical Big Data (BGD), citizen science and open access. The importance of harnessing the vast wealth of historical data alongside modern data sets cannot be overemphasised. Recorded weather data dating back to the 17th century unlocks the potential of historical data for contemporary applications. The data comes in imagery format and has to be digitised and converted to interpretable attributes. It provides scope for validation with data from other sources – e.g., from ice cores (see Section 1.1: Through these applications the centre seeks to deliver humanities knowledge within Japan and beyond through collaborative initiatives. CODH research aspirations fit nicely with Sustainable Development Goals (SDG) and which lead to Big Data Modelling of Sustainable Development Goals (BDSDG) (Mwitondi, K. and Munyakazi, I. and Gatsheni, B., 2020). CODH's activities align with those of the International Research Centre for Sustainable Development Goals (CBAS, 2024).

Section 2 Recommendations

The report makes a number of general and specific recommendations based on the interactive conversations the author had with scientists and researchers at different research centres. The recommendations hinge on how ROIS-DS can optimally utilise the resources to harness the potential that is currently buried in different data archives at the centre. One obvious example that emerged was the on-going development of the AMIDER database at the PEDSC. The development, still its initial stages, has the potential to address some of the challenges other divisions of the institute faces. For instance, the meteorite storage room currently stores both imagery and chemical composition data in separate filing systems. While the files are appropriately coded for unambiguous identification of each meteorite, matching each image to its chemical composition is a tedious and lengthy process. Apparently, the same applies to the ice cores data. Both data types are crucial in classification of the meteorites and indeed in any other relevant analyses. Getting the files linked together will provide not only easy access, but also a much better way of validating analyses based on either data type.



AMIDER Database: This tool, developed at PEDSC, presents an excellent example of collaboration between ROIS-DS centres. NIPR's imagery and chemical composition data on meteorites and ice cores can be stored within AMIDER for analytical and visualisation purposes, using integrated cloud & IoT technologies.

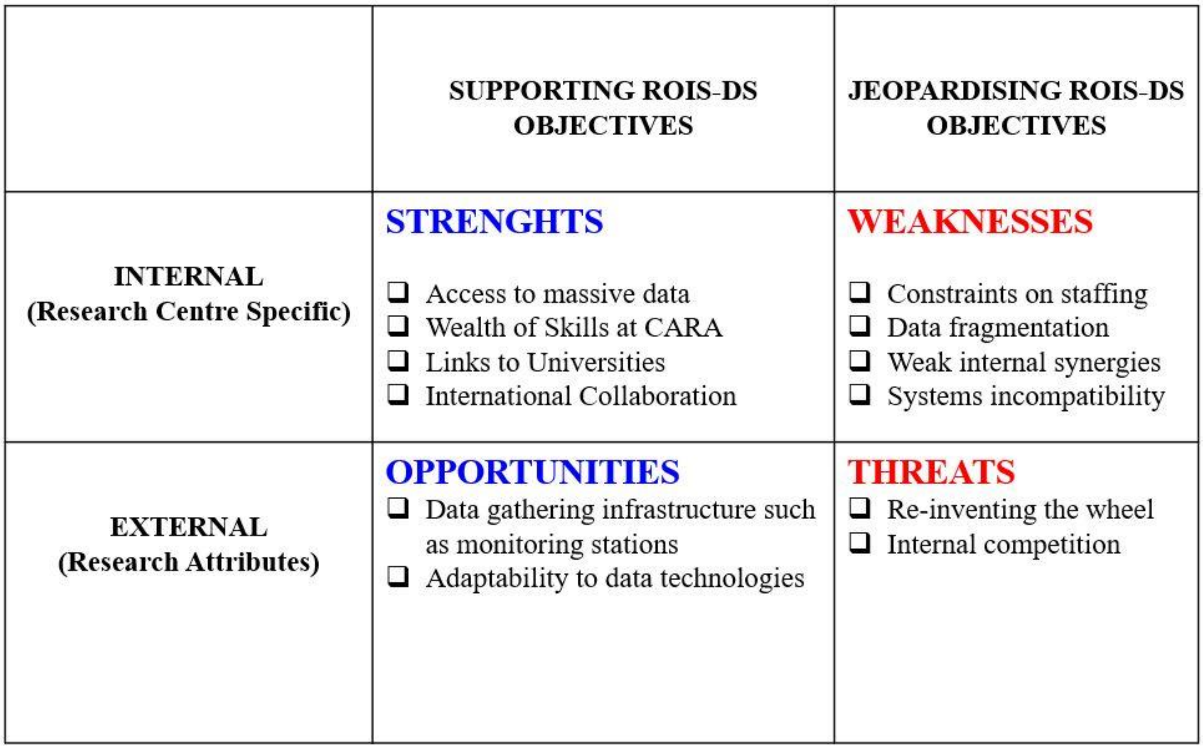
Figure 2: Integrated imagery and chemical composition data for meteorites and ice cores

Figure 2 graphically illustrates integration of data from the meteorites and ice cores centres as part of piloting the AMIDER database. The recommendation is for the AMIDER development to embed a pilot project on only the meteorite centre, as there is already development in that direction. Further, imagery and chemical composition data for meteorites is already available and well-documented, which minimises both back and front-end development efforts. Upscaling the project to include other centres listed in Figure 2, as well as providing access to external collaborators will be much easier once the pilot is successful.

The interactive meetings we had with staff at different research centres revealed that there was unlimited potential for more cross-centre engagement across the institute. In many cases it was noted that there were constraints on staff to carry out what are, sometimes, basic but time-consuming tasks – for instance, mapping thousands of images to their chemical compositions. The report therefore makes specific recommendations as outlined below.

Section 2.1: Identifying Strengths and Opportunities

Timeliness, accuracy and comprehensiveness in uncovering knowledge hidden in the data is what would, typically, determine the usefulness of the data and related project initiatives. However, like in all fields, data generation capacity far outweighs its processing capacity and as such ROIS-DS needs to take a multi-dimensional assessment of the way its centres generate and utilise data. Table 1 provides a synopsis of a standard SWOT analysis.



	SUPPORTING ROIS-DS OBJECTIVES	**JEOPARDISING ROIS-DS OBJECTIVES**
INTERNAL **(Research Centre Specific)**	**STRENGHTS** □ Access to massive data □ Wealth of Skills at CARA □ Links to Universities □ International Collaboration	**WEAKNESSES** □ Constraints on staffing □ Data fragmentation □ Weak internal synergies □ Systems incompatibility
EXTERNAL **(Research Attributes)**	**OPPORTUNITIES** □ Data gathering infrastructure such as monitoring stations □ Adaptability to data technologies	**THREATS** □ Re-inventing the wheel □ Internal competition

Table 1: Selected SWOT analyses aspects of ROIS-DS' data challenges and opportunities

Among the institute's greatest strengths is its access to data – PEDSC being a typical example, wealth of modelling skills at CDARA, links to Japanese Universities and established international collaborations. These strengths tie in nicely with the opportunities they create, such as the ease with which research centres can adapt new data technologies. A typical example are the meteorites and ice cores centres that already have access and continue to amass massive amounts of data from Antarctica. Likewise, near-real-time data from the monitoring station and the seismological observatories present great opportunities.

It is evident, in many cases, that ROIS-DS' well-established data capturing infrastructure can achieve greater results than it does now. There is also the risk of duplication of data repositories and so it is imperative to ensure that key steps are taken to consolidate research data initiatives to avoid internal competition. Figure 3 exhibits a couple of examples of cross-institutional and cross-centre collaborations. Most tools and skills applied at ISM are readily applicable at other institutes, such as NIPR. ROIS-DS stands to benefit from promoting such collaborations. Obviously, such engagement requires planning, managing and funding and can be achieved by addressing a major bottleneck that became evident at each centre – i.e., constrained staffing.

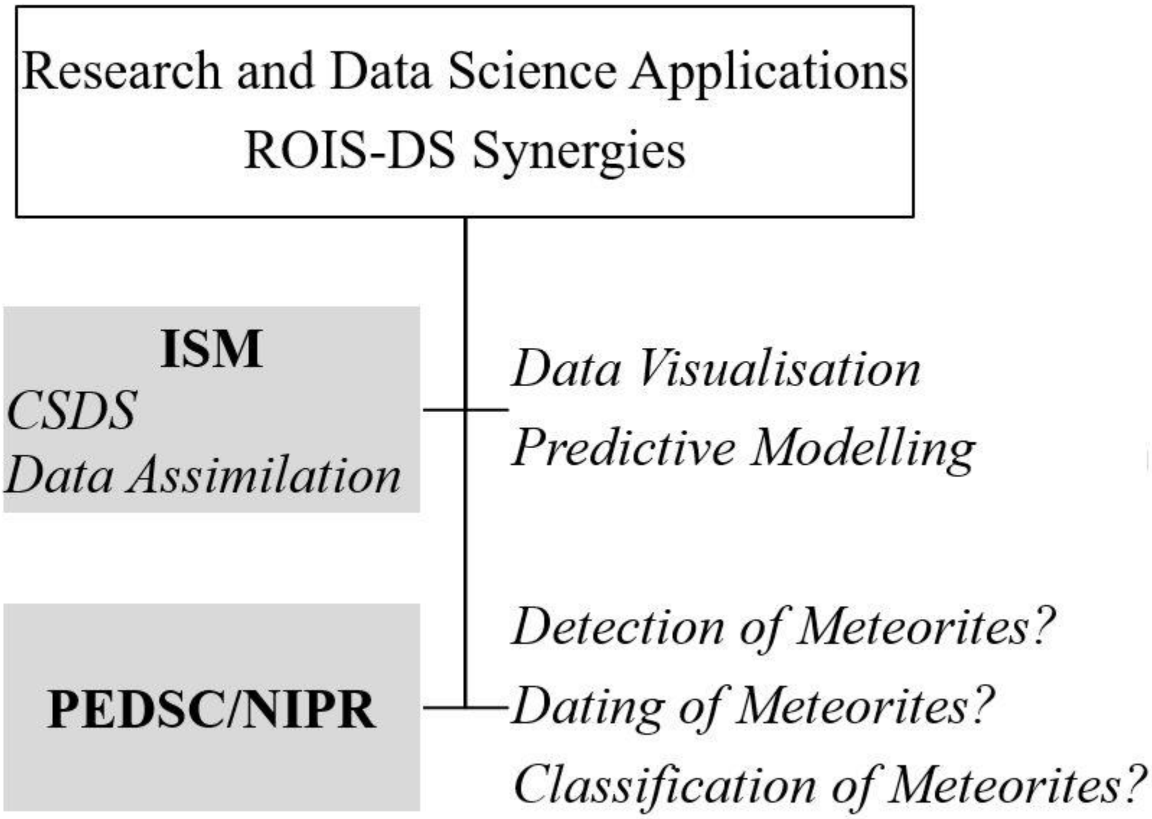


Figure 3: Potential cross-centre data-intensive activities

Clearly, addressing staffing issues can be challenging but the institute can create an interdisciplinary working environment that draws knowledge and skills from its surrounding environment – such as industrial placement, enhancing inter-research centres potential and adapting new technologies to existing data harnessing infrastructure, as summarised below.

Section 2.2: Industrial Placement schemes for students

One way of dealing with data deluge when faced with time-consuming and repetitive tasks, such as mapping thousands of images to their chemical compositions, would be to engage short-term, non-contract staff. Industrial placement schemes are known for providing not only good value for money but also a great way for imparting knowledge on young scientists and researchers. Further, evidence from literature suggests a relationship between knowledge sharing and organizational innovation performance (Zhao, Jiang, & Peng, 2020).

Section 2.3: Cloud-based Data Sharing & Modelling

Using the meteorites example, images captured can instantly be shared with AMIDER via cloud. Again, this will still require manual intervention – but likely to be minimal, as it may only require linking camera tools to cloud and an AMIDER connection to the cloud. Potential future developments of this scheme may include Internet of Things (IoT) applications for sharing images with a host computer and/or cloud. Such applications may also be feasible using IoT sensors fitted to the chemical analysing equipment for instant sharing analyses with a host computer and/or cloud. It must be emphasised that such a process is initially costly, as it may require adaptation to the data gathering tools, but it is quite cost-effective in the long-run, and it will generate massive datasets that are readily available for scientific analysis.

Section 2.4: Harnessing untapped internal potential

Another way of striking a balance between NIPR's data deluge and constrained resources is to tap into internal potential, sharing data, resources and skills. For instance, CDARA has developed several machine learning algorithms for different applications, and these can readily be adapted to applications in other centres. The main challenge here is how to match problem situations with potential solutions. Working culture in individual organisations tend to create silos that are hard to break, unless doors are kept open for sustainable and regular in and out flows of ideas. Table 2 illustrates different approaches that ROIS-DS can adopt to ensure that knowledge gaps and potential solutions across the institute are identified in a timely manner.

Initiative	Description
Monthly cross-centre seminars	Many seminars are already taking place at research-centre levels. This can be enhanced by ▪ Regular/monthly ROIS-DS-wide seminars. ▪ Themes are agreed annually by all centres. ▪ Managed advertised months in advance. ▪ Speakers drawn from the ROIS-DS family.
Interdisciplinary research projects	Cross-centre seminars potentially lead to interdisciplinary research ideas. Interdisciplinary research projects may range from application of research grants to PhD supervision. There is evidence that some of this is already happening.
Cross-centre peer review	A formalised ROIS-DS-wide peer review of grants applications and journal submissions will make it possible for research centres to share vital data and insights into potential collaborations.

Table 2: Tapping into ROIS-DS' internal resources potential.

The initiatives in Table 2 are not exhaustive but they span across the entire spectrum of collaborative interdisciplinary research, and they can be perceived as one major initiative. Managing the initiative across the institute may place further staffing strain on already strained research centres. It is imperative for such a cost to be incurred as it potentially leads to efficiency in utilisation of scarce resources – particularly skills and domain knowledge.

Section 2.5: International Collaboration and Benchmarking

All research centres that were visited exhibited great non-orthogonality in their research activities. There are many examples to illustrate this but suffice it to say that natural and physical phenomena that are investigated at each of the centres aren't happening in isolation. It is therefore imperative for ROIS-DS to see itself in the mirror of some "benchmarking institutions" and to closely collaborate with international institutions that are embarking upon activities that align with, affect or are affected by, its own research activities.

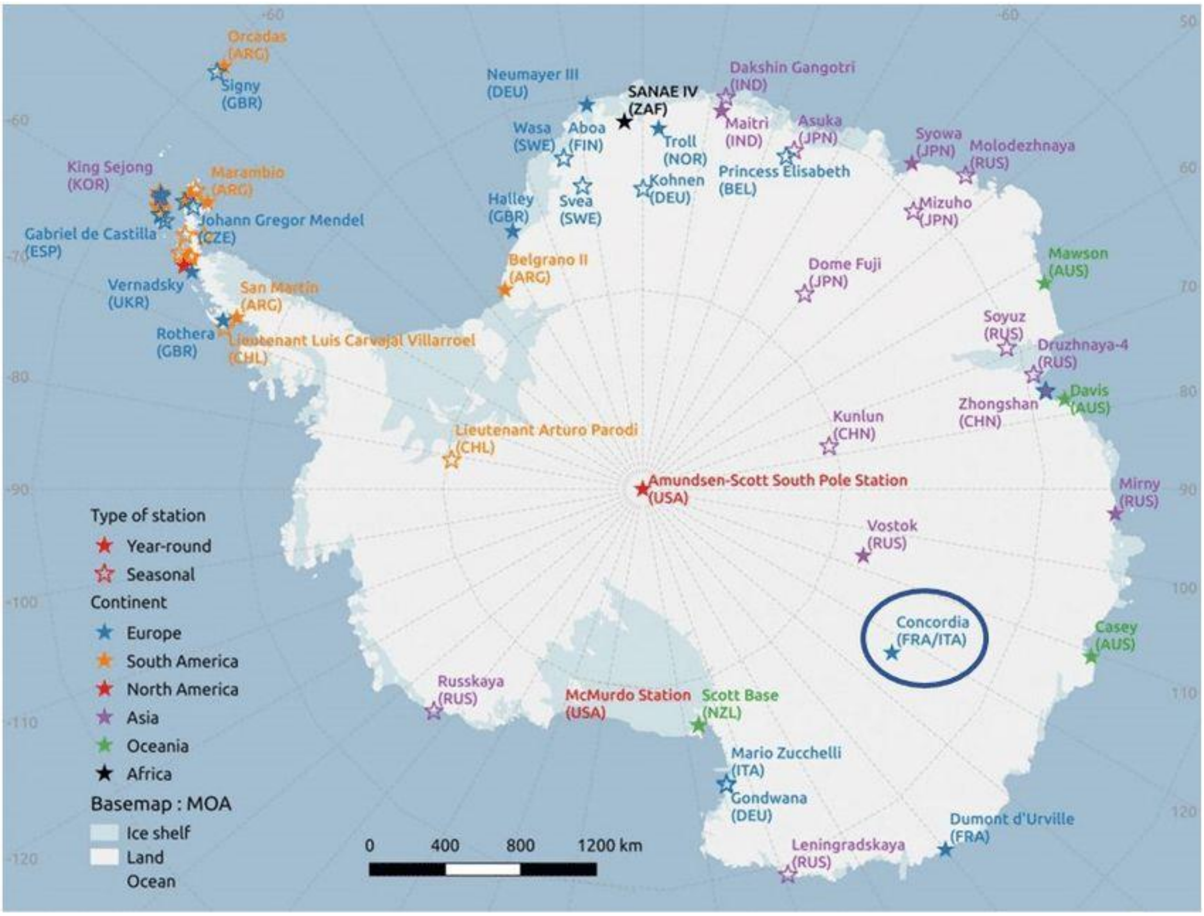


Figure 4: Antarctica monitoring stations by country.

To get an insight into the impact of non-orthogonality in data analysis consider the different locations in Figure 4 (Snels, M. and Colao, F. and Cairo, F. and Shuli, I. and Scoccione, A. and De Muro, M. and Pitts, M. and Poole, L. and Di Liberto, L., 2021) which are managed by different countries. It is reasonable to assume that as the thickness of the ice layers vary with distance from the Amundsen-Scott South Pole Station, say, the chemical compositions of the air bubbles contained in the samples collected may differ. The only way to address these variations would be to validate them against samples collected at other stations. The report proposes the following modes of collaboration based on benchmark institutions and synergies.

Section 2.5.1: Forging Data Intensive Partnerships

Given its position in the ROIS network, PEDSC plays a unifying role when it comes to data generation and consumption among all centres. It is imperative that PEDSC develops sustained partnerships with data-intensive institutions. Partnerships with data practitioners should be designed to enhance existing capacities within the centre and tapping into potentials outside it. There are already established collaborative lines that PEDSC can tap into, including the two affiliates of the International Science Council (ISC, 2024) – i.e., the Committee on Data (CODATA, 2024) and the World Data System (WDS, 2024). PEDSC can significantly enhance its data management, processing and modelling potential by setting up strong a collaboration with the International Technology Office of WDS (WDS-ITO, 2024) and/or WDS data centres.

The ITO has a well-established working relationship with the polar research community that seeks to promote a unified research interface for data generators, consumers and managers across the community. Thus, forging data-intensive partnerships with such institutions will provide PEDSC with the potential not only to latch into novel technologies that are developed and applied across the community, but also to contribute towards the creation of open-source solutions for researchers. One of the infrastructure settings that can be of mutual benefit is facilitating access to polar data even under conditions of poor internet connectivity.

Section 2.5.2: Some Benchmark Institutions

PEDSC can fulfil its unifying role by drawing lessons from institutions that conduct research related to any of the research centres in the ROIS-DS network. One outstanding benchmarking institution is the British Antarctic Survey (BAS, 2024), particularly its affiliate – the UK Polar Data Centre (UKPDC). The UKPDC is United Kingdom's focal point for Arctic and Antarctic environmental data management, and it covers a wide range of interdisciplinary research. Its research and data activities, highlighted in Table 3, align nicely with ROIS-DS centres.

Selected UKPDC Projects	Description	ROIS-DS Relevance
Marine Metadata	Enhancement of availability and accessibility of BAS marine data	ROIS-DS/DBCLS
Polar Airborne Geophysics	Access to airborne survey data	Geophysics apps
Atmospheric Data Access	Data visualisation/access for polar regions' atmospheric & space weather	
Herbarium Collection	Collected dried plant specimens from the Antarctic, sub-Antarctic and surrounding continents	ROIS-DS/DBCLS
Geological Collection	Over 200,000 individual rock and fossil specimens collected from Antarctica and the sub-Antarctic islands and thousands of meters of sediment core from the surrounding seabed	ROIS-DS/PEDSC meteorites research
Bedmap	Collaborative community project for mapping datasets of Antarctic ice thickness and bed topography using a data such as ice-thickness, bathymetry, surface altitude and grounding lines.	Ice cores research at ROIS-DS/PEDSC

Table 3: UKPDC Benchmarking activities for ROIS-DS

The diversity of research activities within the ROIS-DS network entails adaptation of existing data infrastructure to emerging cloud and IoT computing technologies. ROIS-DS will need to carry out a comprehensive study as there are many options (Sharma, B. and Obaidat, M., 2020). The National Centre of Excellence for Food Engineering (NCEFE, 2024) (NCEFE) at Sheffield Hallam University, for instance, adapts novel AI approaches to food engineering. The centre attracts a lot of interest from partners globally, working with industry partners and using interns to fill skills gaps. Apparently, ROIS-DS centres can find adaptations practical and beneficial.

Other relevant benchmarking examples are in the type of research that ROIS-DS centres are embarked upon. For instance, research activities at all centres effectively address one or more of the 17 Sustainable Development Goals (SDG) (DESA, 2023). Attainment of SDGs varies enormously across geographical regions, depending on a wide range of factors, many of which remain unknown (Mwitondi, K. and Munyakazi, I. and Gatsheni, B., 2020). In its quest to enhance its knowledge base and contribute to global scientific achievements, ROIS-DS can draw lessons from research centres with a focus on data science applications ranging from research data validation to cloud-IoT adaptation as shown in selected example in Table *4*.

Institution/Centre/Department	Action/Core Activities	Relevance to ROIS-DS
Data Intensive Research of South Africa (DIRISA)	Data-driven science supports to SANAP	Bi-annual datathons in collaboration with South African Universities
United Nations Statistics Division (UNSTATS)	Big Data applications	Filling skills gaps
International Research Centre of Big Data for Sustainable Development (CBAS)	Monitoring of SDGs	Complements the work of both ROIS-DS and NIPR. Its major project. CBAS' major project - Big Earth Data Science Engineering Program (CASEarth) incorporates Big Data and Cloud Services

Table 4: Selected examples of research institutions that align with ROIS-DS.

The Data Intensive Research of South Africa (DIRISA, 2004) provides data service to the South African National Antarctic Programme. DIRISA acts as the data hub for polar research through the South African National Antarctic Programme (SANAP, 2024) which monitors the natural environment and life in the Antarctic & Southern Ocean via data-driven science. DIRISA also supports space science through the MeerKAT radio telescope Square Kilometre Array (SKA) telescope (SKA, 2024). In collaboration with the National Integrated Cyber Infrastructure System (NICIS, 2024) and the Centre for High Performance Computing (CHPC, 2024), DIRISA plays a key role in grooming young scientists and researchers via sustained bi-annual datathon and hackathon schemes. These schemes not only help groom these youngsters, but they also help fill skills data science and cybersecurity gaps via internships and placements.

With an already established working relationships with Universities across Japan, ROIS-DS symposia and/or conferences could emulate such schemes by engaging students from partnering Universities. However, running such schemes will require establishing a dedicated person/s within PEDSC to organise and manage the activities. Datathon and hackathon schemes are also run by the United Nations Statistics Division (UNSTATS, 2024), focusing mainly on Big Data applications. The institutions in Table 3 and Table 4 provide benchmarking criteria for guiding ROIS-DS' data and general operational processes. Research centres can develop their own metrics for measuring the quality and cost of internal activities for liaison among them. This can be based on internal SWOT to help identify sources of opportunities for improvement and the data activities, which resonate with each centre. PEDSC's AMIDER database is probably best placed to draw lessons from the foregoing benchmarking institutions.

Section 3 Concluding Remarks

This report sought to highlight the potential of ROIS-DS' individual research centres, as part of a coherent research infrastructure. Apparently, ROIS-DS has its short and long-term strategies covering the management and future plans. The report is based on information gathered on short visits to the research centres listed in Section 1. Among the recommendations made include addressing staffing shortfalls in data-related activities – not anymore from the collection point of view than from data management, processing and modelling. Apparently, the ultimate goal of data modelling is application-specific, which entails internal collaborations among the centres and externally with institutions outside Japan.

All the recommendations made require proper and careful planning, executing and monitoring. For instance, all centres typically are embarked upon different projects at any one time and so one way of ensuring that the recommendations are successful would be to carefully evaluate individual project proposals before, during and after implementation. From a data science point of view, any robust monitoring and control will require aligning with the dynamics in the field – that is, adapting to global trends. A few benchmarking examples were identified to sign post ROIS-DS into the direction of good practice in keeping abreast with new developments. It is recommended that ROIS-DS keeps pace with recent developments in Artificial Intelligence (AI) and the related Explainable Artificial Intelligence (XAI) and how these are likely to impact its research centres. For instance, validating numerical analysis using image analysis or vice versa renders itself readily to validation of meteorites chemical compositions with their ML image classification. Such an application is a manifestation of the so-called Multimodal Machine Learning (MML) (Zhang, C. and Yang, Z. and He, X and Deng, L., 2020).

Advances in algorithmic computing and automation of complex processes–like the foregoing mapping of chemical compositions of meteorites to their classified images may require running ML algorithms on small, low-power devices, such as microcontrollers or other IoT devices that transmit data from, say, chemical analysis equipment to the AMIDER server. It is important to avoid quick fixes, when dealing with highly dynamic Big Data. For instance, Generative AI (Hacker, P. and Engel, A. and Mauer, M., 2023) is becoming increasingly popular, as staff with minimal training in data science can carefully utilise to generate realistic content. Much as this may look a blessing, it is also a curse as knowledge gaps in underlying mechanics of data science techniques may potentially lead to undesirable consequences.

Recursive Big Data Modelling (Mwitondi, K. and Storrar, R., 2023) is another trend that may empower algorithms to learn by recursively running through previous paths. The model has the potential to provide robust solutions under highly volatile data conditions or when there are limitations in training data. Hence, PESDC and other ROIS-DS centres need to pay attention to Ethical and Explainable Artificial Intelligence (EXAI) (Izumo, T. and Weng, Y., 2022) to provide fairness, transparency, accountability and trustworthiness. Paying attention to EXAI is crucial, not least because ROIS-DS research data activities span across the entire spectrum of humanity – from citizen science applications–conducted by residents of habitable lowlands in the cold regions, say, to sophisticated research technologies utilising satellites, cloud and IoTs.

References

BAS. (2024). *British Antarctic Survey*. Retrieved from https://www.bas.ac.uk/data/uk-pdc/

Buizert, C. (2021). The Ice Core Gas Age-Ice Age Difference as a Proxy for Surface Temperature. *Geophysical Research Letters, 48*(20).

CBAS. (2024). *International Research Centre for Sustainable Development Goals* . Retrieved from http://www.cbas.ac.cn/en/

CHPC. (2024). *Centre for High Performance Computing*. Retrieved from https://www.chpc.ac.za/

CODATA. (2024). *Committee on Data*. Retrieved from https://codata.org/.

DESA, U. (2023). *The Sustainable Development Goals Report 2023.* New York: SUN DESA.

DIRISA. (2004). *Data Intensive Research Initiative of South Africa* . Retrieved from https://www.dirisa.ac.za/

Hacker, P. and Engel, A. and Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* New York: ACM Digital Library.

ISC. (2024). *International Science Council*. Retrieved from https://council.science/

Izumo, T. and Weng, Y. (2022). Coarse ethics: How to ethically assess explainable artificial intelligence. *AI and Ethics, 2*(3), 449-461.

Ludescher, J. and Bunde, A. and Franzke, C. (2016). Long-term persistence enhances uncertainty about anthropogenic warming of Antarctica. *Climate Dynamics, 46*, 263–271.

McLachlan and Peel. (2000). *Finite Mixture Models.* New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics Section.

Mwitondi, K. and Munyakazi, I. and Gatsheni, B. (2020). A robust machine learning approach to SDG data segmentation. *Big Data, 7*(97).

Mwitondi, K. and Storrar, R. (2023). A Recursive Method for Identifying Glacial Landforms in Hillshaded Areas Using Digital Elevation Data and CNN Models. *DSWS 2023 International Symposium on Data Science 2023.* Tokyo: ROIS-DS.

NCEFE. (2024). *National Centre of Excellence for Food Engineering*. Retrieved from https://www.shu.ac.uk/national-centre-of-excellence-for-food-engineering

NICIS. (2024). *National Integrated Cyber Infrastructure System*. Retrieved from https://www.nicis.ac.za/

ROIS-DS. (2016). *Joint Support-Centre for Data Science Research*. Retrieved from Research Organisation of Information Systems: https://ds.rois.ac.jp/en_aboutus/en_notice/

SANAP. (2024). *South African National Antarctic Programme*. Retrieved from https://www.sanap.ac.za/

Sharma, B. and Obaidat, M. (2020). Comparative analysis of IoT based products, technology and integration of IoT with cloud computing. *IET Networks, 9*(2), 43-47.

SKA. (2024). *South African MeerKAT (Square Kilometre Array - SKA) Telescope* . Retrieved from https://www.sarao.ac.za/gallery/meerkat/

Snels, M. and Colao, F. and Cairo, F. and Shuli, I. and Scoccione, A. and De Muro, M. and Pitts, M. and Poole, L. and Di Liberto, L. (2021). Quasi-coincident observations of polar stratospheric clouds by ground-based lidar and CALIOP at Concordia (Dome C, Antarctica) from 2014 to 2018. *Atmospheric Chemistry and Physics*.

Stewart, R., & Benepe, O. &. (1965). Formal Planning: the Staff Planner's Role at Start up. *California: Stanford Research Institute*.

UNSTATS. (2024). *United Nations Statistics Division*. Retrieved from https://unstats.un.org/bigdata/index.cshtml

WDS. (2024). *World Data System*. Retrieved from WDS-ISC: https://worlddatasystem.org/

WDS-ITO. (2024). *International Technology Office*. Retrieved from https://wds-ito.org/

Zhang et. al. (2020). Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: A review. *Climate Dynamics, 54*, 5127-5144.

Zhang, C. and Yang, Z. and He, X and Deng, L. (2020). Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing, 14*(3), 478-493.

Zhao, S., Jiang, Y., & Peng, X. &. (2020). Knowledge sharing direction and innovation performance in organizations: Do absorptive capacity and individual creativity matter? *European Journal of Innovation Management, 24*(2).