Digital Antarctica

D. Summary and Recommendations

June 2022

Australian Antarctic Program Partnership
Institute for Marine and Antarctic Studies
20 Castray Esplanade, Battery Point
Tasmania, 7004, Australia
aappartnership.org.au

AAPP
Private Bag 80,
Hobart Tasmania,
7001

Version information

Version	Description	Author	Date
1.0	Release Version	Rob Jennings	11/07/2022

Executive Summary

Digital Antarctica is an Australian Antarctic Program Partnership (AAPP) project which has been running since July 2020 with the aim to define a framework within which Australian Antarctic data can be shared. This document summarises the project, its progress, and recommendations for the future. The appendices of this document summarise key elements of the *Digital Antarctica* document suite.

A fully realised *Digital Antarctica* will consist of standards, designs, and services. Standards define the requirements that a service must meet to be *Digital Antarctica* compatible and will be used to create designs for individual services. Those designs will, in turn, be used to create the services which will make up *Digital Antarctica*. Once created, those services will be built, tested, and implemented at the data centres across the AAPP. Any person, application, or service wishing to access Antarctica data will be able to access it via the *Digital Antarctica* services, which will deliver data in a manner which meets the *Digital Antarctica* standards, regardless of its source.

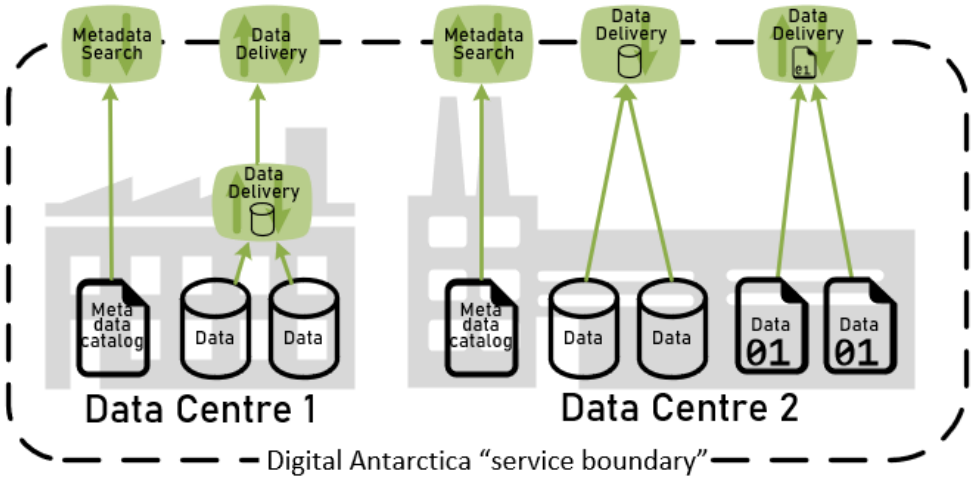

Figure 1 - The Digital Antarctica service boundary

Data shared via *Digital Antarctica* services will be logically grouped to facilitate creating both the standards and the services. A dataset may be grouped by any combination of its research categories, its data format, the type of instrument that captures it, the collections it belongs in, the services that currently serve it, or any other logical grouping. This will ensure that standards are broad enough to be practical, but narrow enough to define distinctive and useful differences between the groups.

When beginning design and implementation of *Digital Antarctica,* the project should start small, finding criteria for early successes which demonstrate ability, while ensuring any solution is scalable and directly leads to further larger-scale success. Sharing previously under-utilised datasets could demonstrate high value while lessening risk through negative impact. Standards could also be developed from common services, such as existing OGC services. Work should also continue on the machine learning demonstrated in the virtual database prototype.

At all stages, stakeholder engagement is key to the success of *Digital Antarctica*, and should be constant and consistent, to ensure needs are met and work is not only endorsed but championed by stakeholders.

The AAPP is winding down the *Digital Antarctica* project and handing over responsibility for the delivery of a shared data framework to the AAD's Integrated Digital East Antarctica program, which began in 2022. The scope of the IDEA project has not been fully explored or defined, but its mission to "facilitate and coordinate the acquisition, analysis and synthesis of Antarctic and Southern Ocean data" aligns with the *Digital Antarctica* goals. All *Digital Antarctica* documentation and recommendations will be available to IDEA, as will the Digital Antarctica Reference Group.

Digital Antarctica

Overview

Digital Antarctica is an agreed standardised framework to facilitate data sharing across multiple Antarctic research organisations in way that aligns with the FAIR data principles, which advocate that data should be Findable, Accessible, Interoperable and Reusable.[1]

Each data centre within the AAPP stores data in a number of different formats, and shares data via various mechanisms including services, portals, and direct access to files. They may also have one or more services to search and retrieve metadata, separate from their data delivery services. As each centre has its own history, its own data, and its own users, the methods used for sharing data vary over the data centres. They also have established processes and workflows.

Digital Antarctica will provide standards for services that share Antarctic data. Any tool looking to access Australian Antarctic data will be able to connect to services built to these standards to search and retrieve data in a way that is consistent across the *Digital Antarctica* data centres.

Figure 2 - The green boxes with the yellow ticks represent Digital Antarctica compliant services which serve data requests in a standard manner.

The standards will define input and output parameters for the services, ensuring that data are requested and provided in a consistent manner across all data centres. These standards can be applied to existing services or to new services. New services may be built to replace existing services or to adapt data from existing services so that the data are presented in a manner that meets the *Digital Antarctica* standards.

Due to differences in data types and data groups, data centres will require multiple *Digital Antarctica* compliant services to serve all of their data.[2] The exact number of services will depend on the grouping, and on the data that the data centre serves.

[1] *See https://www.go-fair.org/fair-principles/ for more details on FAIR*
[2] *See* Data grouping *for more information on data types and groups*

Example

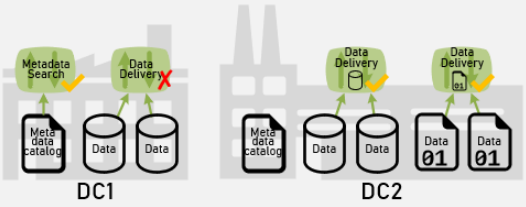

Figure 3 - Data Centres with missing or non-standard services

A data centre (DC1) currently has 1 service for metadata search and retrieval, and another to serve its database contents.

A second data centre (DC2) has no service for metadata search and retrieval, but has a service to serve its database contents, and another service to serve datasets stored in its file storage.

DC1's metadata service meets *Digital Antarctica* standards; however, its database service does not. DC2's current services all meet *Digital Antarctica* standards; however, it does not currently have a metadata search and retrieval service.

For DC1 to be *Digital Antarctica* compliant, a new database adapter service must be built to a) take the standardised *Digital Antarctica* inputs and adapt them to the inputs of the existing service and b) take outputs of its existing service and adapt them to the standard *Digital Antarctica* outputs.

For DC2 to be *Digital Antarctica* compliant it will need to build a new metadata service to *Digital Antarctica* standards.

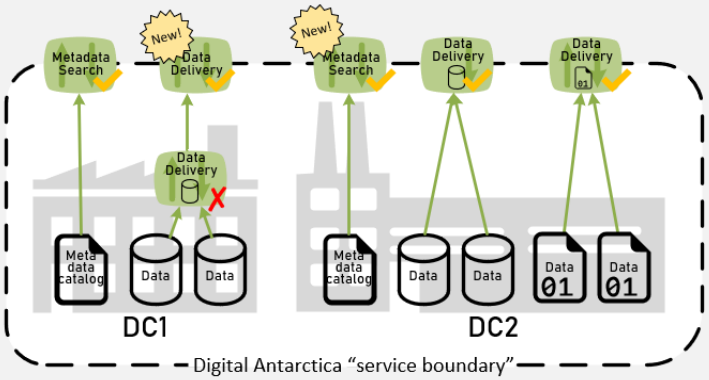

Figure 4 - DC 1 has a new adapter service for data, and DC2 has a new metadata service. All compliant services form a service boundary, which serves all data to Digital Antarctica

Any system wanting to access any *Digital Antarctica* enabled data will be able to use consistent service architecture to access the services, and will retrieve data from those services in a consistent format, regardless of the data centre it is accessing.

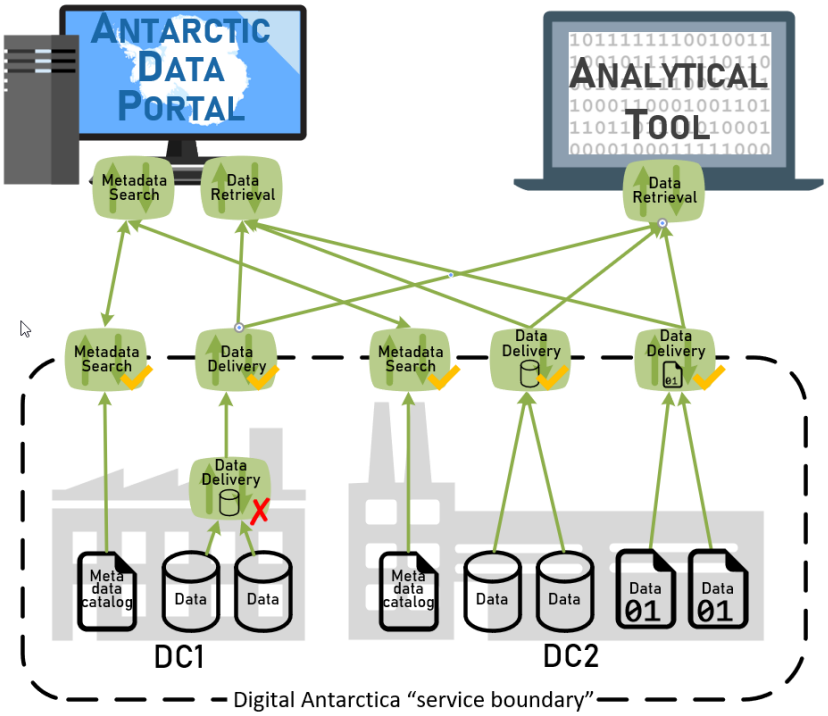

Figure 5 - A portal might use a metadata search service to search multiple Digital Antarctica data centres, whereas an analytical tool might connect directly to data

Objectives

A primary goal for *Digital Antarctica*, extrapolated from the 2017 review of Australian Antarctic Science Program Governance by Drew Clarke[3], is to enable "a world-class centre for Antarctic data analytics", which demonstrates that Australia is a leader in Antarctic data, supporting Australia's Antarctic Treaty obligations and bolstering its position in Antarctic negotiations. This strong focus on data will help Australia not only meet its Antarctic Treaty obligations, but exceed them by displaying a pro-active use of data to meet policy and research needs and thereby demonstrating respect and understanding of the region.

This goal will be reached by achieving the objectives listed below. The low-level objectives are the actual deliverables of the *Digital Antarctica* program[4]. Having those low-level objectives delivered will lead to achieving the mid-level objectives, and the mid-level objectives will in turn lead to the high-level objectives.

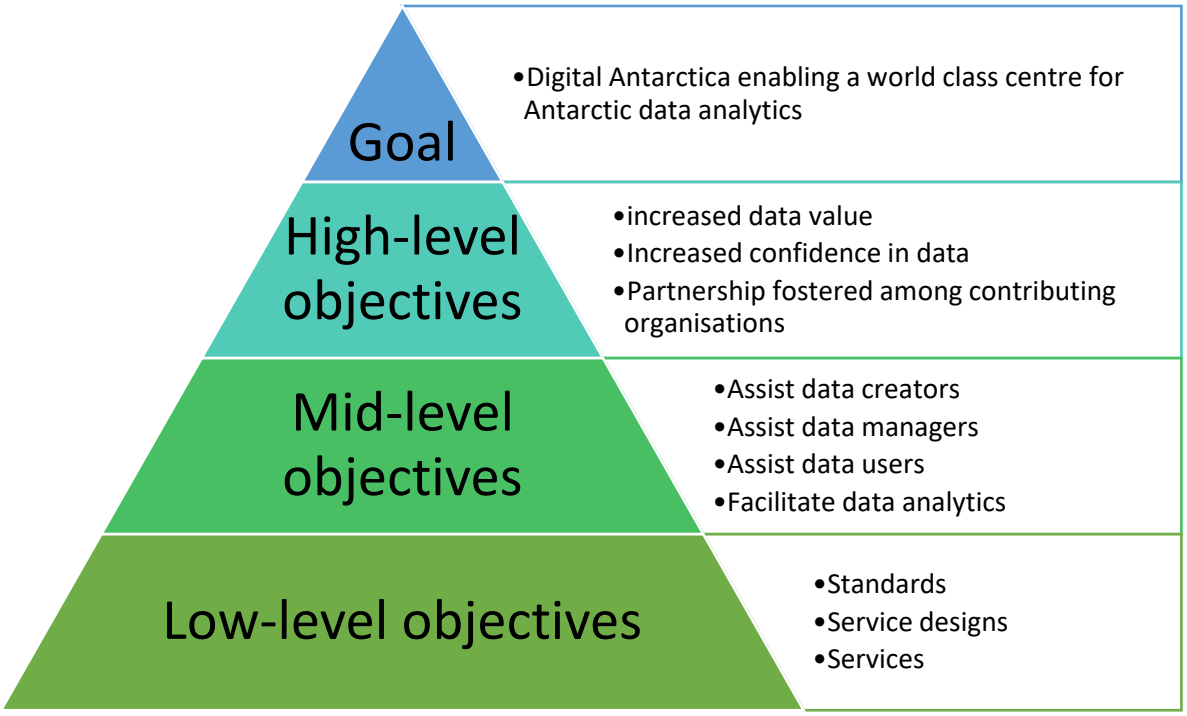

Figure 6 - Hierarchy of objectives. See each below for further explanation of each level

Goal

Digital Antarctica enabling a world class centre for Antarctic data analytics.

High-level objectives

- Increased data value – Because of its remoteness, Antarctic data are expensive to collect. A comprehensive data model which enables data consumers to bring together science from decades of research across a number of organisations will distribute the value of each of those collections across multiple use cases.
- Increased confidence in data – The strong standards employed to serve data via *Digital Antarctica* will ensure that a data consumer will have confidence in the completeness, and the provenance, of those data.

[3] Clarke, D. 2017. *The Australian Antarctic Science Program Governance Review.* https://www.dcceew.gov.au/science-research/antarctic-review

[4] *See* Creation and implementation of Digital Antarctica *Creation and implementation of Digital Antarctica for more information on deliverables.*

- Foster partnership among contributing organisations – Through continual consultation and engagement, contributing organisations will work together to ensure that *Digital Antarctica* meets not only its overall goals, but also their specific needs and the needs of their users.

Mid-level objectives

- Assist data users[5] to better store, manage, share, find, and use data, by ensuring their data are Findable, Accessible, Interoperable and Reusable (FAIR). The following data users will see benefits from this assistance:
  - *Data creators* (anyone who generates research data – e.g. scientists):
    - by ensuring their data are accessible and interoperable.
  - *Data managers* (anyone involved in the curation and maintenance of data – e.g. anyone who facilitates the upload of data, ensures the quality of data and metadata, and who maintains hardware and services used in capturing serving and otherwise sharing data)*:*
    - by providing ongoing guidance and standards for best practice in capturing, storing, and sharing data.
  - *Data consumers* (anyone who uses data that have been shared – e.g. researchers, government departments, policy makers and advisors, educators):
    - by ensuring that data from multiple sources are easy to find and obtain
    - by ensuring that data are standardised and interoperable, meaning the data from multiple sources can be assimilated with ease
- Facilitate data analytics by integrating with wider datasets, which will improve the breadth and effectiveness of the data products for end users.

Low-level objectives

- A suite of documented standards.
- A suite of service designs using those standards, to facilitate:
  - Metadata search
  - Standardised metadata delivery
  - Standardised data delivery
  - Data interoperability
- Services built to those designs (in conjunction with bespoke services for each data centre to interact with those standardised services).

The low-level objectives map directly to the *Digital Antarctica* deliverables. See *Creation and implementation of Digital Antarctica* below for more information on deliverables.

Additional benefits

Antarctic data stored in *Digital Antarctica* compliant data centres will be more easily ingested into tools designed to find and deliver metadata. While *Digital Antarctica* may not provide a search portal, it will deliver data and metadata in such a way that existing search tools, or any new portals, will be able to easily use them.

By providing standardised data inputs and outputs and providing frameworks which foster interoperability, *Digital Antarctica* will enable analytical tools to directly access data from multiple sources, ready to be analysed in situ, or via online based services. Tools such as Jupyter Notebooks will be able to directly access data, and can be shared and expanded upon when more data are required, without significant re-tooling of the code which accesses the data.

By delivering consistent data, *Digital Antarctica* will present Australian Antarctica data as a cohesive and unified data landscape. While each data centre will retain autonomous control over their data and

[5] *See Appendix 5c – Data Users for more information on Data users.*

their remit, the services they use to present their data will ensure that *Digital Antarctica* is seen as an authoritative source of Australian Antarctic data.

Creation and implementation of Digital Antarctica

Deliverables

A fully realised *Digital Antarctica* delivery will have documented standards, service and interface designs, and functions (APIs, web services, applications, user interfaces etc.) built to those designs. The standards and design documents will be the core documentation suite for *Digital Antarctica*, and will be referenced by any organisation wishing to make their data available via *Digital Antarctica*.

To initially create the core documentation, each required standard and design document must be identified. This will be documented in a *Standards and service register*, which will help with project planning, but also act as a register for the ongoing documentation suite.

Digital Antarctica may be built iteratively, with the delivery of standards, designs and functions spread across a number of consecutive or concurrent projects. However, each function requires a design and each design will require one or more applicable standards. For example, if a *Digital Antarctica* prototype was to handle only ocean CTD data presented as netCDF files, the standards for the keywords and vocabularies of the CTD data and the standards of the services to share the netCDF files would need to be established; and the services to share the data and the functions to call those services would all need to be designed and built.

Standards and service register

Each service in *Digital Antarctica* will serve a set of data in a manner that best suits the data and the service. This will be determined by factors such as the type of research, the format of the data, the requirements of users of those data, and the capabilities of the service (see Data grouping for more details). The sets of data served by a single service may be broad or narrow, depending on the service and the data. By deploying a combination of *Digital Antarctica* services, a data organisation will be able to share all of its in-scope Antarctic data via *Digital Antarctica*.

Each service, and each standard that is used to design that service, must be identified. The service and standards register will identify each required service and standard, with a description of the service or standard, its purpose and its scope.

The register may be a document, but may also be a task board or ticketing service or other similar tracking system. Regardless, it will aid the initial *Digital Antarctica* project in the creation of the standards and services, and will also aid any organisation wishing to make their data available via *Digital Antarctica*.

Standards

The standards delivered as part of *Digital Antarctica* will form the core of the *Digital Antarctica* framework. Any functionality created for *Digital Antarctica* will be a tangible output of those standards, and any organisation wishing to make their data available via *Digital Antarctica* should begin with the standards. The exact number of standards will depend on a large number of factors, the most significant of which is how data are grouped (see Data grouping).

Standards will be defined for the services that deliver data as well for the data and metadata being delivered.

Service standards will describe the service, and will include technical details of the service. They will also detail the conditions that apply for its use and its standard inputs and outputs. This documentation will also include describe standards required to enable interoperability with recognised analytical tools (such as Jupyter notebooks).

Metadata standards will include required elements to be recorded with the metadata (for example, keywords and spatial information), as well as data schema information (for example vocabulary integration, parameter information and data structure information). These standards may be an extension or profile of existing metadata standards, or may be something new.

Keyword standardisation will form part of the metadata standards. This will include documentation of standardised keywords, vocabularies, and parameters for all services.

Many service and metadata standards currently exist, and it is not the goal of *Digital Antarctica* to create new standards where an applicable one already exists. Where an existing applicable standard exists, that standard may be used in formation of the *Digital Antarctica* standard. The existing standard may be used completely, in part, or not at all, depending on the exact scope and definition. The documented *Digital Antarctica* standard may re-iterate an existing scope and definition or may defer to it entirely. In all cases, any existing applicable standard should be acknowledged.

Any standards will undergo a stakeholder approval process to ensure they meet stakeholder needs.

Design

For each required function, *Digital Antarctica* will produce a design. A function is anything that needs to be created to enable *Digital Antarctica*. It will most likely be a piece of software, and may be a user function (such as a web page or application) or a system function (such as a web service or system log). It may be a new function, or may involve updating an existing function (including reconfiguring an existing system).

For a function which is to be built specifically for *Digital Antarctica*, the design will contain all required detail to describe and build the function. For a function that already exists (e.g. a previously built function, or off-the-shelf product) that will be adapted to use with *Digital Antarctica*, the design will specify the nature of its use and any modification or configuration requirements.

As with standards, designs will undergo a stakeholder and technical approval process to ensure that they will meet stakeholder needs while also being technically feasible.

Function

Once designed, a function would be built and deployed. The exact process to build and deploy a function will be determined by the software development methodology, the project, and the organisation in question, however may involve one or more of the following steps:

- Build – This is usually performed by one or more appropriate developers, using the designs. Depending on the function, this may be undertaken in one or more development environments (on a developer's computer, on an organisation's network, in the cloud, or any combination thereof).
- Test – There are multiple levels of testing that may be applied. Testing may be performed manually or using automated tools. Examples of testing include:
    - Developers, as part of the build process, will perform formal and ad-hoc tests to ensure the function performs at a basic level.
    - System testers can perform further tests to ensure a function's operation, but also test that a function will meet the business requirements.
    - If a user interface is being created, User Experience testing may be undertaken to ensure that the function is intuitive and meets user expectations.
    - Stakeholder, or users on the behalf of stakeholders, may also perform user acceptance testing, to ensure that the function performs as expected before agreeing to deploy the system.
- Implementation – This is the process of deploying the function into a production environment, and may include acceptance testing and sign-off.

Data grouping

As discussed in Deliverables above, data may be grouped to facilitate sharing. This grouping would not be a physical grouping (as it is not the intention of *Digital Antarctica* to store or replicate any data), but rather would be used to find commonality between disparate datasets with a goal of finding logical ways to share and consolidate similar data.

There are a number of dimensions upon which data may be grouped for this purpose. These dimensions could include the field of research that the data encompass, the filetypes of the data, or the way the data are shared. A single dataset will fit into groups within any of these dimensions, and depending on the data may fit into many groups within a single dimension.

Field of research

As most of the data shared via *Digital Antarctica* would be scientific data, they can be grouped by the field or fields of research that the data encompass. Many of the AAPP data centres already use NASA's GCMD Earth Science keywords[6] to help classify and enrich their metadata.

There are 14 high-level topics under the *Earth Science* category, and each topic has multiple underlying sub-topics (Term, Variable_Level_1, Variable_Level_2, Variable_Level_3 and Detailed_Variable). This creates over 3000 possible unique topic/variable combinations that can be generated and applied to a dataset. The data centres across the AAPP currently use approximately 1000 of these unique combinations.

The 14 top-level Earth Science GCMD topics are:

- Agriculture
- Atmosphere
- Biological classification
- Biosphere
- Climate indicators
- Cryosphere
- Human dimensions
- Land surface
- Oceans
- Paleoclimate
- Solid earth
- Spectral/engineering
- Sun-earth interactions
- Terrestrial hydrosphere

There are a number of factors to consider when looking at the GCMD keywords as a grouping dimension:

- These categories are broad and, at their highest level, could not be used to usefully group data for the purposes of sharing. Oceans, for example, contains all variables related to oceans from acoustics to chemistry to ocean winds.
- Similarly, using all 3000 unique variables would not create logical groupings as those variables become too granular.
- Some topics may have many logical groups within them
- Many datasets have more than one GCMD Keyword associated with them (for example, the AAD datasets have, on average, over 4 unique GCMD Keywords per dataset)

[6] https://www.earthdata.nasa.gov/learn/find-data/idn/gcmd-keywords

- Some sub-variables may have more in common with each other than with their parent topic. For example, *Oceans>Ocean Chemistry* may be more readily grouped with *Terrestrial Hydrosphere>Water Quality/Water Chemistry* data than with other data under the *Oceans* topic.

Data format

Another method of grouping data is by their file and sharing formats. All data available via *Digital Antarctica* must be digitally available and shared. How the data are stored and shared could help define its grouping.

Different formats of data include:

- Database/tabular – datasets which are stored in tabular format, including spreadsheets, delimited text files (such as CSVs) and database tables.
- Document/Text – Any file that stores data as human readable text, either formatted or unformatted. This would include, html, pdf, rtf, and txt files.
- Spatial – Any file that is designed to be used and viewed natively in spatial tools such as a GIS or NetCDF viewer. This would include NetCDF, SHP, WFS, and GRID data formats.
- Video/image – Any file that displays a visual image, either still or moving. This would include AVI, BMP, JPG, and MOV formats. Note that some collections of still images are shared as video files.
- Audio – Any audio recording, including WAV and MP3 file formats.
- Proprietary/other – Certain file types are only accessible by the software packages that created them. Other files are not immediately recognisable by their filetypes. While these files may fit in to one of the above categories, they cannot be automatically shared with them without intervention.

As well as these file formats there are also other ways to group data by type. For example, the data feature class (e.g. point, profile, time series, trajectory, grid). These would, for the most part, be sub-categories of the spatial data above, but may also be present in other data formats, and could help provide granularity when using format as a grouping mechanism.

Instrument/instrument category/collection

Each instrument that collects data will have its own set of parameters that it collects. Similar instruments, or instruments that collect similar types of data, may be grouped as like instruments. Likewise, instruments that use similar vocabularies for their parameters may be grouped. These might include:

- Drones
- Argos floats
- Ship instrumentation

Similarly, data collected as a group, such as underway data or CTD data, may be logically grouped together.

Existing services and sharing methods

A practical way to group data for sharing is by services that are currently used, or are currently available, to share data. Many data centres, for example, share WMS, WCS, and WFS data via their own GeoServer facilities. Spatial data are often made available via THREDDS servers. Many of these sharing facilities are relatively standard across data centres, and so may provide a simple first step in grouping data.

Intersections

The above represents ways in which data can be grouped along a single dimension. However, when looking at the best ways to share data, groups from multiple dimensions may be intersected with each other.

Using the *Data format* dimension as an example, there might not be a single best way to share database data. But when intersecting that dimension with the *Field of research* dimension, we may find that chemistry data stored in a database might be best shared one way, while marine biological data stored in a database might be best shared another way.

Expanding this concept to encompass intersections of all dimensions may help in finding the most logical groupings of data.

Priorities

Having established the logical groupings, the *Digital Antarctica* project must then decide which group or groups to address first, or how otherwise to proceed. This will be determined by considering the following factors.

Impact

Impact represents the change an implementation would trigger.

Impact based on service changes

Implementing a *Digital Antarctica* service standard will cause a varying degree of change depending on the differences between an existing service and the new service standard. *Digital Antarctica* may, for example, re-use an existing in-use service standard for a particular data group. This would mean that there would be no implementation required for data centres already using services built to that standard. This would be a potentially low impact change.

Conversely, *Digital Antarctica* may introduce a service standard and service design for a data group that requires an existing service to be updated. Implementation in this scenario would likely be high impact, as the alteration may have an effect on current user interactions with the service.

Impact based on data usage

The volume of traffic a particular data group experiences will also have an impact.

Implementing even a small change on datasets that have high volumes of traffic will create a high impact, particularly if that traffic is to multiple users or user groups.

Implementing a change on datasets that are currently under-utilised, and in particular are currently under-served may have a large beneficial impact while carrying a relatively low risk. There is, however, no guarantee that these datasets would be simple (see **Error! Reference source not found. Error! Reference source not found.**).

Impact based on data size

Making any changes to services that serve data that is not necessarily high traffic, but large in size, may also have a significant impact.

Consideration of factors

Having identified high and low impact activities, the following factors should also be considered.

- *User requirements* – Input from key user groups defining key questions, and pressing problems that need addressing
- *Knowledge gathering* – Certain deliverables may yield useful learnings that the team can apply in future work
- *Complexity, feasibility, and effort* – The complexity of designing, building, and implementing a solution for a specific data group will vary depending on the data group, the capabilities of the team and resources available at each data centre.
- *Resources* – The resources available to the *Digital Antarctica* team, e.g. in funding, staff, and stakeholder availability
- *External drivers* – e.g. political influence, stakeholder availability, scope creep

These factors, combined with impacts, provide a framework for decision making.

Site-specific activities

While *Digital Antarctica* will specify standards and designs for general use, each data centre will have its own analysis and development to perform. Data centres will need to determine:

- what data should be made available via *Digital Antarctica;*
- the readiness of those data to be served via Digital Antarctica;
- which services currently meet *Digital Antarctica* standards;
- the best strategy for serving data that doesn't currently meet the Digital Antarctica standards. This may include data uplift, service uplift, or creation of new services;
- services which require uplift to meet *Digital Antarctica* standards;
- which new services need to be built, either as new services, or as adapter services

The design, build and implementation of site-specific services will be the responsibility of the data centres, with the assistance of the *Digital Antarctica* program.

Recommendations

During the June 2022 Digital Antarctica Reference Group meeting, the group discussed what should be considered when starting to create and implement a framework like *Digital Antarctica*. A number of recurring themes emerged from that discussion:

- A multi-faceted approach
- Demonstrate early success
- Ramp up – demonstrate ability at a small scale, but with scalability/future proofing built in
- Demonstrate end-to-end feasibility.
- Recognise value in smaller/under-utilised datasets
- Address stakeholder requirements

These themes are explored further below.

Early success, ramp-up and end-to-end feasibility

Early success can show stakeholders and interested parties that the proposed solution is feasible. However, if not approached carefully, it can come at a cost of scalability. While it is important to demonstrate functionality early, that functionality must be reflective of the whole proposed solution, and be future-proofed for scalability. A small solution that relies heavily on manual data manipulation, for example, will not easily apply to large datasets, and so does not meet the requirement of scalability.

This early success should also demonstrate its potential for full end-to-end functionality. For example, if it is a small step of what will be a series of larger steps, then how it fits into that larger picture should be clear and demonstrable. If, on the other hand, it is an end-to-end solution of a thin slice of data, then it should show how it will handle larger data volumes.

Once that ability is demonstrated at a small scale, a logical path to gradually increase scale should be defined and agreed upon.

Under-utilised datasets

When considering impacts and looking for early signs of success, it is recommended to identify datasets that are currently under-utilised. By finding value in these data, and making them accessible in ways they have not previously been, *Digital Antarctica* will demonstrate its worth while not impacting negatively on existing usage.

It is important to consider that the datasets, even smaller ones, may be under-utilised due to some level of complexity. But this in itself may then be an excellent source for knowledge gathering and learning.

Stakeholder engagement

Stakeholder engagement and support is important to project success, and this engagement should be consistent throughout the project.

Consultation with stakeholders, including data users, such as the research community and policy makers (see Appendix 5c – Data Users for more examples), will help the project by answering questions as well as creating networks of support. These groups currently have data needs, data processes, and data holdings. Consultation will help discover what they can bring to *Digital Antarctica* to bolster it, as well as provide information on how *Digital Antarctica* can help them address their needs. Moreover, this consultation will create engagement and championship in the project, which will contribute to its success.

Digital Antarctica also requires direct ongoing participation from data managers. Consultation is one approach to ensure data managers are engaged, but further commitment will be required from managers of the data to ensure *Digital Antarctica's* success. Investigation should be undertaken to determine how best to incentivise ongoing participation in the upkeep of the *Digital Antarctica* standards and services.

Pathway to implementation

When looking at the data groups and recommendations above, a multi-faceted approach is preferred. The project is encouraged to identify strong elements from multiple options and work on those simultaneously, either in combination or in parallel.

Specifically:

- Find common OGC services among the contributing organisations. These are existing services, currently serving similar data across multiple data centres. The effort required to standardise these already similar services, and to document those standards as *Digital Antarctica* standards, will be minimal but will provide an early success.
- Continue exploring and developing the machine learning analysis and Virtual Database technology (see Appendix 4 – CAST Collaboration, below). This will enable at-scale data and metadata analysis that will, in turn, help determine data groups and to unify disparate datasets

Integrated Digital East Antarctica

Australian Antarctic Division review

In 2021 the document "Leading Australian Antarctic Science – Review of Australian Antarctic Division Science Branch" (the O'Kane review) was published[7]. It outlines a number of recommendations for the whole Division, including its focus on science, its decadal plan, and its infrastructure. It also includes a specific recommendation on creating an "Integrated Digital East Antarctica" (IDEA). The document mentions *Digital Antarctica* as "consistent with the Australian Antarctic Science Strategic Plan", and as "platform for a new digital initiative", and goes on to recommend building IDEA with a primary goal to "create, manage and maintain a digital twin of East Antarctica and surrounding Southern Ocean". This goal aligns with *Digital Antarctica*, but potentially has a much larger scope in terms of delivery and stakeholders.

[7] https://www.antarctica.gov.au/science/information-for-scientists/changes-to-the-australian-antarctic-science-program/okane-review/

The Antarctic Division has agreed with all recommendations in the O'Kane review, and has begun work to address the recommendations, including preliminary work on the IDEA program.

AAPP

Digital Antarctica came from recommendations in the "Australian Antarctic Science Program Governance Review" (the Clarke Review)[8], published in 2017, which called for "a comprehensive data model of the Australian Antarctic Territory." The AAPP management committee has recognised that the *Digital Antarctica* project has come to its funding conclusion, and has agreed to transfer responsibility of the digital data model raised in the Clarke review to the AAD and IDEA. The AAPP recognises that the goals of *Digital Antarctica* and IDEA align in a way that satisfies the recommendations of the Clarke review.

Pathway to IDEA

While the IDEA program is likely to build upon the foundations that *Digital Antarctica* has laid down, the full program scope and direction have not yet been decided. However, the IDEA team have developed the following mission statement:

> *The Integrated Digital East Antarctica (IDEA) Program will facilitate and coordinate the acquisition, analysis and synthesis of Antarctic and Southern Ocean data. IDEA will bring together a broad suite of stakeholders to lead the cutting-edge development of Antarctic simulation and modelling science to answer the most pressing scientific questions facing Antarctic and the Southern Ocean.*

This aligns strongly with the *Digital Antarctica* vision:

> *To support Australia's Antarctic treaty interests and obligations by defining an interoperable, comprehensive, and sustainable approach to data sharing.*

It also aligns with the *Digital Antarctica* scope statement, which says:

> *Digital Antarctica will facilitate access to all publishable Australian Antarctic and Southern Ocean research data and relevant ancillary data stored by the AAPP partner organisations.*

Until the IDEA program is fully launched, there is no directly defined path from *Digital Antarctica* to IDEA. The early stages of the IDEA program are likely to further define its parameters and scope, and in doing so will recognise the *Digital Antarctica* work and expand on how that work will be used moving forward.

[8] https://www.dcceew.gov.au/science-research/antarctic-review

Appendices

Appendix 1 – Glossary

Term	Description
AAD	The Australian Antarctic Division
AAPP	The Australian Antarctic Program Partnership. A partnership of Australian Antarctic research organisations with the goal of better understanding the role of the Antarctic Region. The partnership includes the following partner agencies: • University of Tasmania • Institute for Marine & Antarctic Studies • The Australian Antarctic Division • CSIRO • Bureau of Meteorology • Geoscience Australia Tasmanian Government
BoM	Bureau of Meteorology
CTD	Conductivity, Temperature and Depth. A collection of data containing these observations.
DARG	The Digital Antarctica Reference Group, a group of participants representing the partner organisations of the AAPP, with the purpose of discussing and driving the *Digital Antarctica* project.
FAIR	FAIR is an acronym that describes attributes of data in terms of shareability. The acronym stands for: • *Findable* – This attribute defines how easily the data can be found based on their metadata. Data that are richly described and tagged, and that have unique identifiers (such as DOIs) are considered findable. • *Accessible* – This attribute defines how easily the data can be accessed, based on where and how they are shared. For data to be accessible they must be able to be retrieved by both humans and machines • *Interoperable* – This attribute describes how well the data can be integrated with other data and data centres. • *Reusable* – This attribute describes how ready a dataset is to be re-used or repurposed. This includes determining how applicable the data are outside their own initial purpose, as well as their provenance and how attributable they are.
Function	Any piece of software that is used within *Digital Antarctica* This could include services, applications, web interfaces.
GA	Geoscience Australia
GCMD	Global Change Master Directory. NASA's international data collection resource. Now available via https://idn.ceos.org/index.html The GCMD also hosts a repository of keywords that can be applied to metadata to help categorise it. A user interface to browse those keywords is available here: https://gcmd.earthdata.nasa.gov/KeywordViewer/scheme/all?gtm_scheme=all
IDEA	Integrated Digital East Antarctica – a new data initiative starting at the AAD
IMAS	Institute for Marine and Antarctic Studies
IMOS	Integrated Marine Observing System
Jupyter Notebook	A web based interactive data science and scientific computational environment.
NASA	The National Aeronautics and Space Administration, an independent agency of the U.S. federal government responsible for the civilian space program, as well as aeronautics and space research.
netCDF	Network Common Data Form – a form of array-oriented scientific data.

Service	The word “service” has a number of real-world definitions, usually regarding an amenity or facility that is performed for someone (e.g. a cleaning service or a ride-sharing service). However, in terms of systems and data delivery (and in terms of *Digital Antarctica*), the term “service” refers to a piece of software that exposes and delivers data or functions from a system to an external source. A web service hosted by a data centre, for example, allows a person or system to access some of that data centre’s data via the web without being granted access to the data centre’s whole systems.
The Clarke Review	The Australian Antarctic Science Program Governance Review, published in 2017, by Drew Clarke. Available at: https://www.dcceew.gov.au/science-research/antarctic-review
THREDDS	Thematic Real-time Environmental Distributed Data Service – a service that provides human and machine access to data files, including netCDF files.
User interface	Any user facing tool used to access data within *Digital Antarctica*. This could be a web page, an application on a computer or phone, or some other tool.
User story	A user story is a software development tool that describes a user, a task that user wishes to perform, and a reason that the user wants to perform the task. They are usually written in the format of “As a [user or type of user], I want to [an action that the user would like to perform] so that [a goal that performing that action will achieve].” User stories highlight the various users of a tool, and the benefits that the tool provides them. They can be used during development as a measure of progress or success, and also help to give a personal perspective to user requirements.
UTas	The University of Tasmania

Appendix 2 – References

Clarke, D. 2017. *The Australian Antarctic Science Program Governance Review.* https://www.dcceew.gov.au/sites/default/files/env/pages/7753423c-a411-480e-b1d8-8669a098d33d/files/aus-antarctic-science-program-governance-review.pdf

O’Kane, M. 2021. *Leading Australian Antarctic Science - Review of Australian Antarctic Division Science Branch.* https://www.antarctica.gov.au/site/assets/files/57785/aad_science_branch_review_report.pdf

Appendix 3 – The Digital Antarctica Reference Group

In late 2020, the Digital Antarctica Reference Group (DARG) was created from representatives of the AAPP partner organisations, to align stakeholder organisations in the definition and progress of *Digital Antarctica*.

At the time of its creation, *Digital Antarctica’s* point of interest was specifically the data involved in Antarctic research. As such its members were representatives of the AAPP who are the data experts in their organisation, and who had the required knowledge to inform discussion. While the DARG started with data managers, the vision of the group was that it would evolve as the project continued, to reflect the needs of the project at the time.

The primary purpose of the reference group is as a communication channel. The group acts as a single access point for the project to communicate with the AAPP members and to seek advice and expertise when required. It also acts as a forum for members to discuss related activities with each other. Each of the *Digital Antarctica* documents has been created in consultation with, and with the review of, the members of DARG.

Over the course of the engagement, there have been numerous online meetings with the entire group, and with individual representatives of the group. On occasion, interested parties from outside

the group have presented to, and attended, the group meetings. While participation has been consistent, there have been a number of representative changes over the years reflecting the staffing changes at the partner organisations. Despite this, there has almost always been a representative from each partner organisation in the reference group.

In June 2022, the first face-to-face DARG workshop was held at IMAS in Salamanca, Hobart, with representatives from most of the partner organisations attending in person or online. This event also hosted representatives from SAEF, the ARDC and the AAD's IDEA program.

Appendix 4 – CAST Collaboration

One goal of the *Digital Antarctica* project was to investigate a prototype. This goal was addressed via collaboration with UTAS as part of the Centre for Antarctic and Southern Ocean Technology (CAST)[9] initiative. This initiative is an agreement between the AAD, CSIRO and UTas to collaborate on, among other areas, technology and innovation projects.

The *Digital Antarctica* project engaged with Byeong Kang of UTAS and his team who started work on a virtual database which would help with the analysis of data and metadata, and assist researchers in uploading their data and searching for other data.

As a first step, the team built a prototype data upload tool which assists contributing researchers in selecting GCMD Science Keywords for their metadata. While this tool is not directly related to *Digital Antarctica*, the prototype uses machine learning concepts and functionality which analyses data and metadata to identify similar datasets, which would help *Digital Antarctica* in regards to the data grouping concepts and may be used as an underlying and ongoing analysis tool.

Appendix 5 – The Digital Antarctica Document Suite

Over the course of the *Digital Antarctica* pre-analysis phase, the project has produced a number of documents describing the current state, scope, and requirements of the project. These are:

- A1. High-level Scope – A high-level understanding of the scope, based on early discussions with the Digital Antarctica Reference Group. This was to capture the stakeholder understanding of what *Digital Antarctica* could be.
- B1. High-level Current State – A high-level snapshot of what each of the partner organisations do, and how they manage data.
- A2. Refined Scope – A deeper understanding of the scope, delving into some of the concepts touched on in the high-level document. This document also defined the vision and scope statements.
- B2. Refined Current State – Having established the scope of *Digital Antarctica*, this document takes a deeper look at what each data centre does with in-scope data. It also collates those data to provide an overall picture of the Antarctic data landscape
- C. Requirements – A list of requirements that a fully realised *Digital Antarctica* solution should meet.

The remaining appendices provide a basic overview of key concepts covered in the previous *Digital Antarctica* documents.

Appendix 5a – Current State

Each organisation within the AAPP houses data which reflect the specialisation of the organisation in question. For example, the Bureau of Meteorology mainly houses climate data while IMAS mainly houses ocean data. The entire partnership holds data that span most of the GCMD Earth Science categories, but when viewed together, certain patterns do emerge. Most organisations host ocean

[9] https://www.antarctica.gov.au/news/2021/new-centre-for-antarctic-and-southern-ocean-technology/

data, and so across the partnership the AAPP has a notably high instance of ocean data. There are also significant numbers of atmosphere, biosphere, climate, and biological datasets.

Likewise, data formats vary across the organisations and the research types however patterns can be found. A large amount of data in the AAPP are stored in some form of table (e.g. a spreadsheet, CSV, or database), or as some form of spatial data format (e.g. netCDF).

For more details on the data in the below figures, see *Digital Antarctica – B2. Refined Current State*.

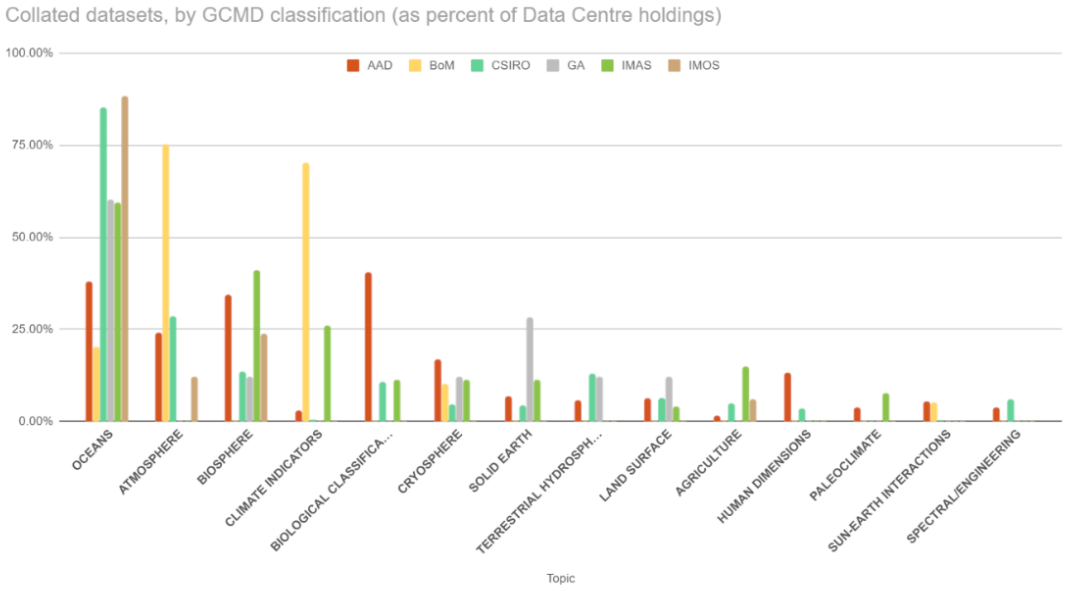

Figure 7 – In-scope datasets by GCMD Topic, expressed as percentage representation within each Data centre (e.g. 38% of AAD datasets have the OCEAN topic).

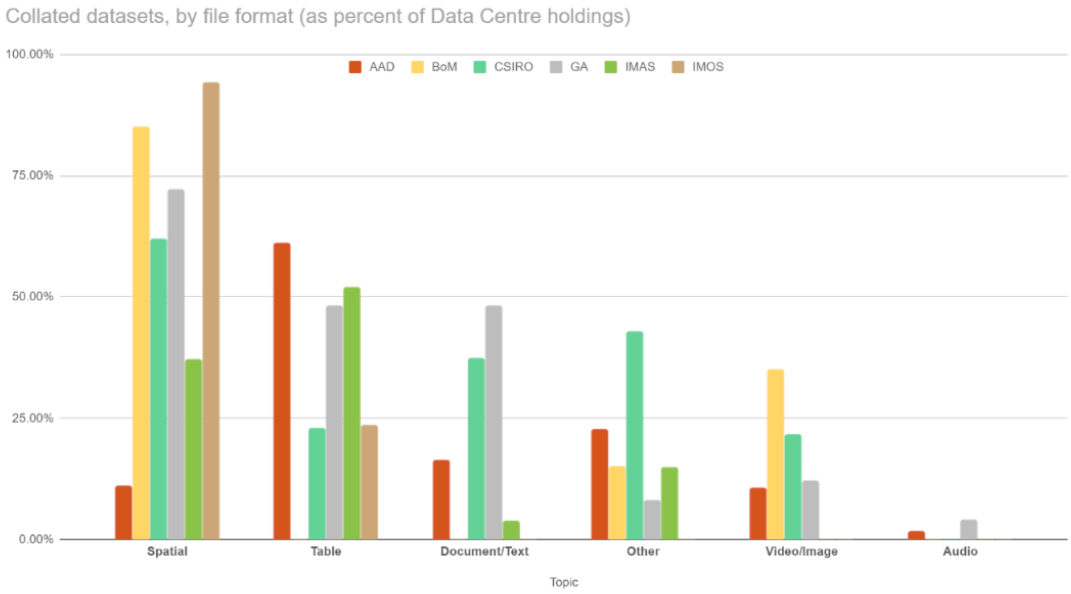

Figure 8 - In-scope datasets by file format, expressed as percentage.

Topic	AAD	BoM	CSIRO	GA	IMAS	IMOS	Grand total
Table	1072	0	141	12	14	4	1243
Other	396	3	265	2	4	0	670

Spatial	194	17	384	18	10	16	639
Document/Text	284	0	231	12	1	0	528
Video/Image	186	7	133	3	0	0	329
Audio	28	0	0	1	0	0	29

Table 1 - In-scope datasets containing data of a particular file type category - expressed as number. Table sorted by total number of datasets across all data centres, per file format category

Appendix 5b – Scope

Digital Antarctica will facilitate access to all publishable Australian Antarctic and Southern Ocean research data and relevant ancillary data stored by the AAPP partner organisations.

As the Antarctic Treaty is one of the drivers of *Digital Antarctica*, it will focus on publicly accessible data. It will include in its scope all such data created by Australian research organisations which focus on, or are created in, the Australian Antarctic Territory; the Southern Ocean; and the regions that impact, or are impacted by, these regions. It will include any data that could be used for research purposes (such as raw data, calibrated data, analysed data, data products, models, and model outputs) and ancillary data (such as code, configuration files and specifications, and supporting documentation).

Scope is further explored in the *Digital Antarctica A2. Refined Scope* document.

Appendix 5c – Data Users

For the purposes of *Digital Antarctica,* a data user is anyone that interacts with in-scope data. The data users of *Digital Antarctica* can be broken up into three broad categories: data creators, data managers, and data consumers.

Data creator

Data creators generate data that will go on to be made available to data consumers. A data creator may be a person, generally a researcher, or may be a system or instrument. The data may be anything from raw observational data to synthesised data products.

Data generated by data creators may often be stored in a data creator's own file stores before being submitted or transferred into a data management store. Data may also be non-digital at this stage, waiting to be automatically or manually digitised.

Data creators may include:

- Researchers (who collect, synthesise, analyse, and create data and data models);
- Data officers who manage instrument data before submitting them to a data centre; and
- instruments which directly collect and share data.

Data manager

A *data manager* is anyone involved in the curation and maintenance of data. This may include anyone who facilitates the upload of data, defines data sharing processes and principles, ensures the quality of data and metadata, and/or maintains software, hardware, and services used in capturing serving and otherwise sharing data, including:

- Data managers
- Data officers
- Instrument technicians
- IT development and support staff

Data consumer

A *data consumer* is anyone who accesses and uses data that have been shared, regardless of their purpose for doing so. This includes:

- Researchers
- Members of the public
- Policy makers and advisors
- Research planners

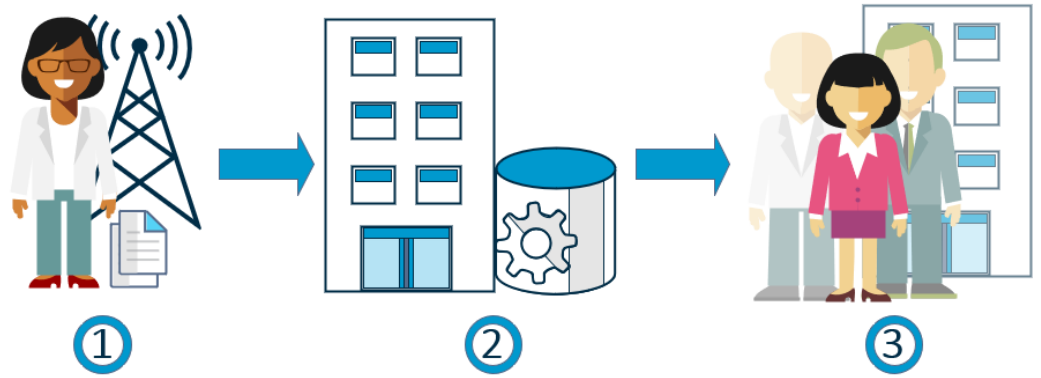

Figure 9 - simple data flow from data creators to data users, via data managers within a data centre.

In this diagram a data creator (1) generates data which are submitted to a data centre (2). The data centre will have one or more data managers who are responsible for ensuring the quality of the data, their metadata, and/or their maintenance. The data manager may also be responsible for ensuring the data are shared appropriately to the data users (3).

Appendix 5d – User Stories

The following user stories show how data served via *Digital Antarctica* can provide benefits across the Antarctic research, policy, and data community. Most of these user stories are further explored in the *Digital Antarctica A2. Refined Scope* document.

Data User type	User Story
Data consumer	As a researcher, I want to supplement my own data with other data that have already been collected so that my research can be well supported.
	As a researcher, I want to incorporate data from multiple sources easily so that I can concentrate on my research rather than on data manipulation.
	As a research planner, I want to see what data have already been collected so that I can plan my observations around the gaps in existing understanding.
	As a policy maker, I want to find data that support the decisions I make, so that policy can be backed by empirical evidence.
	As a member of the general public, I want to be able to find Antarctic and Southern Ocean data and information so that I can use them in my school, work, or area of general interest
Data creator	As a researcher, I want to ensure that my data are findable, accessible, and reusable by as many people as possible so that they can help others and so that interest in this field is maintained.
	As a researcher, I want guidance on the best way of recording my data, so that they are easily sharable.

Data manager	As a data manager, I want to ensure that my data meets interoperability standards so that they are easily accessible and re-usable by data consumers.
	As a data manager, I want access to a community that provides guidance and real-world experience on services, formats, vocabularies, and other data sharing best practices so that my practice can be informed and up-to-date.

Appendix 5e – Requirements

These requirements are further defined in the *Digital Antarctica - C. Requirements* document

Req. ID	Requirement
Functional Requirements	
FR1	*Digital Antarctica* shall enable the access of data across multiple data sources.
FR2	Similar data made available via *Digital Antarctica* shall be presented in a standardised format.
FR2.1	Similar data shall be presented in a standardised structure.
FR2.2	*Digital Antarctica* shall enable standardised nomenclature for similar values.
FR2.3	*Digital Antarctica* shall enable standardised formats for similar values.
FR3	*Digital Antarctica* shall provide full provenance details for any data.
Non-functional Requirements	
NR1	*Digital Antarctica* shall enable interoperability of data across multiple sources.
NR1.1	*Digital Antarctica* shall enable data analytics to be performed using data from multiple data sources.
NR1.2	*Digital Antarctica* shall enable search of data across multiple data sources.
NR2	*Digital Antarctica* shall provide guidance on standards.
NR3	Where practicable, *Digital Antarctica* shall investigate and re-use existing standards and services.
NR4	*Digital Antarctica* data items shall be presented as recorded, without being altered from its source.