Alignment of Polar Data Policies - Recommended Principles

Draft

Contributors: Stein Tronstad (1), Pip Bricher (2), Johnathan Kool (3), Peter Pulsifer (4), Anton van de Putte (5), Jan Rene Larsen (6), Helen Peat (7), David Rayner (8), Taco de Bruin (9), Marten Tacoma (9), Jen Thomas (10), Frank Nitsche (11), VS Samy (12), Dariusz Ignatiuk (13), Erik Buch (14), Anne Treasure (15)

(1) Norwegian Polar Institute, (2) Southern Ocean Observing System, (3) Australian Antarctic Data Centre, (4) Carleton University, (5) Royal Belgian Institute of Natural Sciences, (6) Arctic Monitoring and Assessment Programme, (7) British Antarctic Survey, (8) Swedish National Data Service, University of Gothenburg, (9) Royal Netherlands Institute for Sea Research, (10) Swiss Polar Institute, (11) Lamont-Doherty Earth Observatory of Columbia University, (12) National Centre for Polar and Ocean Research, Goa, India, (13) Svalbard Integrated Arctic Earth Observing System, (14) European Global Ocean Observing System, (15) South African Environmental Observation Network (SAEON) Information Systems

The purpose of this document is to present a basis for alignment of polar data policies, notably the policies and statements of the Scientific Committee on Antarctic Research (SCAR), the International Arctic Science Committee (IASC), the Sustaining Arctic Observing Networks (SAON) initiative, and the Southern Ocean Observing System (SOOS). The document examines the state and recent developments of global and important regional data policies, as well as technological and institutional developments that should or might be considered when forming new polar data policies. Based on this examination we conclude by identifying a number of data management principles that can be regarded as essential to the management of polar research data, and can be incorporated in all polar data policies in such a way that they are aligned with each other and with overarching global and regional data policies. A final part of the document presents an additional set of principles that may or should be included in polar data policies under given circumstances but not universally (such as the CARE principles).

Intended audience

This document aims to inform and support expert groups and science managers involved in revision and continued development of the data policies and statements of IASC, SAON, SCAR, SOOS, and other polar science communities. It is our hope that the document will also be useful to other polar research programmes, polar data centres and data managers, relevant funding agencies, and other experts interested in sound, long-term management of polar data.

Process and involvement

At the Polar Data Forum III, held in Helsinki, Finland, in November 2019, members of all three polar data committees gathered to discuss the rationale for better alignment of polar data policies, investigate recent developments in data-driven polar research, identify core elements of new, aligned polar data policies, and identify which organisations, projects, and people the policy should apply to.

The discussions in Helsinki formed the starting point for this document, which has been further developed throughout 2020-2021, in four sessions of the Polar To Global Online Interoperability and Data Sharing Workshop/Hackathon series, which allowed additional refinement of the policy recommendations presented here.

Part 1, Background and Objectives

Why polar data policies?

Data policies are important tools to set expectations among the science community and other rights holders and stakeholders about how and what data to share and how to treat data shared by others. As a primary resource for science and scientific collaboration data should be managed according to widely recognised principles. A common data policy will clarify obligations and stipulate norms with respect to data sharing, access, management, preservation, and acknowledgment. Agreement on such principles will facilitate collaborative research.

This document focuses on policy for the management of data created through scientific observing and research based in research institutions[1]. However, it is also relevant to data and information generated through other activities and knowledge systems, including operational research and monitoring, Indigenous Knowledge and data, Non-Governmental Organizations, commercial operations, and policy bodies. Important issues relating to Indigenous Knowledge will be further explored later in this document, and a complementary analysis is being conceived that would further develop these topics in collaboration with key knowledge holders and organizing bodies (e.g. Arctic Indigenous Peoples representative organizations). Important progress in this regard for the Arctic region is expected through the SAON ROADS process.

For science coordinating bodies, funding agencies, research institutions, and scientists themselves, good data sharing policies and practices optimise the societal benefits and the scientific utility of the data they collect. Through the transformative effects of digital technologies, data have become increasingly important resources not only for scientific research, but for economic development, environmental protection, resource management and human welfare. In this lies an increasing impetus towards open data, and the now widely accepted assertion that assets generated by publicly funded research should be managed in a way that maximises the public benefit. At the same time a stronger emphasis on transparency and reproducibility in science means that scientific journals increasingly require that all data supporting scientific papers be made openly available. Funding agencies tend to have similar requirements.

These arguments apply even more strongly to polar research, which tends to be physically challenging, often constrained by logistical resources, and extremely costly. Such restrictions increase the value of maximising the utility and reusability of data.

Data collected in polar areas may also have societal benefits for both local and global residents on issues as diverse as natural hazard alerts, resource management, and global ocean and climate monitoring. Between the potential social utility of these datasets and the difficulty in obtaining them, there is particular need for strong data management policies in polar regions.

An additional reason for developing a data policy at the polar level is that much research is conducted as part of interdisciplinary national polar research programs and through international collaborations that are coordinated geographically, rather than by discipline, at the scale of the Arctic, Antarctic, or combined polar regions. It is therefore valuable to have

[1] For a definition of 'science' we will refer to UNESCO's Recommendation on Science and Scientific Researchers: "the word "science" signifies the enterprise whereby humankind, acting individually or in small or large groups, makes an organized attempt, by means of the objective study of observed phenomena and its validation through sharing of findings and data and through peer review, to discover and master the chain of causalities, relations or interactions; brings together in a coordinated form subsystems of knowledge by means of systematic reflection and conceptualization; and thereby furnishes itself with the opportunity of using, to its own advantage, understanding of the processes and phenomena occurring in nature and society".

an aligned data policy that provides common standards across national, institutional, and disciplinary boundaries to support a common approach to data management and sharing.

The development of polar data policies

Polar data sharing and open data policies go back to the First International Polar Year (1882-1883), and data from this ground-breaking international effort remain available today[2].
The international collaboration during the International Geophysical Year in 1957-1958 led to the signing of the Antarctic Treaty in 1959, where one of the fundamental articles of the Antarctic Treaty states that "Scientific observations and results from Antarctica shall be exchanged and made freely available". With this, polar data sharing became a matter of international law.

In the digital age, international scientific organisations started to introduce explicit data policies and data management recommendations, some of which will be presented below. As a brainchild of the International Council for Science (ICSU) and the World Meteorological Organization (WMO), the fourth International Polar Year (IPY, 2007-2009) provided a major impetus to improving data management at both poles and introduced a seminal data policy specific to polar research.

Following the IPY, individual data policies were developed by the polar science groups. In 2010 the Scientific Committee on Antarctic Research (SCAR) adopted a SCAR data policy prepared by its Standing Committee on Antarctic Data Management (SCADM) and largely built on the IPY Data Policy. The International Arctic Science Committee (IASC) followed suit with its Statement of Principles and Practices for Arctic Data Management, or the IASC data statement, in 2013, and established its Arctic Data Committee (ADC) together with SAON the following year. Finally, the Southern Ocean Observing System (SOOS) Data Management Sub-Committee established a similar data policy in 2015.

With the shared IPY pedigree, these three polar data committees - SCADM, SOOS DMSC, and ADC - have developed data policies that are similar. However, while they share major ideas and obligations, they were not written to be explicitly aligned and differ in important aspects. In addition, they pre-date the widespread (FAIR) and emergent (CARE, TRUST) adoption of three key sets of principles for data management:

- FAIR (Findable, Accessible, Interoperable, Reusable) Principles (Wilkinson et al. 2016) that encourage machine-interoperability of datasets,
- TRUST (Transparency, Responsibility, User Community, and Sustainability and Technology) Principles for trustworthy data repositories (Lin et al. 2020)
- CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) Principles for management of data about and collected by Indigenous people.

[2] https://www.pmel.noaa.gov/arctic-zone/ipy-1/Data-P1.htm

Alongside the FAIR, CARE, and TRUST principles, there has been parallel development in data technologies as new sensor technologies are developed and it becomes increasingly feasible to develop and create big datasets (Science International, 2015[3]). A renewed data policy should cover some of the issues associated with big data, data integration, and new sensing technologies.

In recent years, the three polar data committees have worked increasingly collaboratively on a range of issues, including policy discussions, semantics, and federated search tools for metadata records. This increased collaboration and coordination between the three groups is an additional reason for an aligned data policy.

Definitions

'Data' has been fundamentally described as the material basis for transmission of information to humans. Many different definitions exist and may vary by context. In this document we will understand the terms broadly and relate to definitions of 'data' and 'information' that are being developed by the ADC-IARPC-SCADM Vocabularies and Semantics Working Group.

Individual communities may want to limit or expand the definition as appropriate, e.g. "...data generated under the auspices of a [community name]-sponsored research project".

Data

Data: A set of values, symbols, or signs (recorded on any type of medium) that represent one or more properties of an entity. For example, the numbers generated by a sensor, values derived from a model or analysis, text entered into a survey, or the raw text of a document.

Generally speaking, data are used to quantitatively or qualitatively describe one or more persons or objects. Research data provide the evidence base for supporting or refuting ideas in a scientific manner.

Information: Products derived from data that lead to a greater understanding of an entity. For example, (i) the interpretation of a range of data from an array of conductivity sensors across the Arctic Ocean that informs us about that ocean's salinity range or (ii) the narrative text of a report on harmful algal blooms that informs the reader on the timing of these blooms.

[3] Science International (2015): Open Data in a Big Data World. Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP)

Metadata

Metadata is information that describes the data source and the time, place, and conditions under which the data were created ('data about data'). Metadata informs the user of who, when, what, where, why, and how data were generated. Metadata allows the data to be traced to a known origin and known quality.

Metadata can be used for discovery and identification of data collections; to provide information on structural aspects of the data, and to provide administrative information on aspects such as ownership and licensing.

Part 2, Reference Policies and Policy Drivers

As the institutional framework for international scientific collaboration evolves, so do the data policies of the global organisations. In the following sections we will briefly present the current data policies of selected global and regional organisations that have been leading the way towards more sustainable management of the knowledge assets represented by scientific data. Our purpose here is to identify the core principles, fundamental recommendations, and best practices emerging from overarching, global and regional data policies that the polar data policies should be aligned with or take into consideration. The selection of policy documents gives priority to guidance from UN bodies, and from the International Science Council (ISC) with affiliates, the Group on Earth Observations (GEO), and the Organisation for Economic Co-operation and Development (OECD) because of their particular roles in science and scientific data management.

Reference policies

Global organisations

International Science Council

The International Science Council (ISC, formerly ICSU and ISSC) is the parent body of IASC and SCAR. By itself and through its subsidiaries, CODATA (the ISC Committee on Data) and WDS (the ISC World Data System), the ISC has made several important data policy recommendations.

In its Assessment on Scientific Data and Information (2004) ICSU observes that "science has long been best served by a system of minimal constraints on the availability of data and

information", and that a strong public domain for scientific data and information promotes greater return from investment in research, stimulates innovation and enables more informed decision-making. Thus, one of the fundamental recommendations of the assessment is that "ICSU should continue to actively promote the principle of full and open access to scientific data". A comprehensive set of further recommendations are offered, inter alia on interoperability (32), long-term accessibility (38), sound management of IPR (39-40), ensuring data integrity (29-31), professional data and information management (16-21), and the use of metadata (22-25).

The World Data System (WDS) Data Sharing Principles (2015) is a more recent document guiding the activities of the WDS, which is an interdisciplinary body of the ISC that certifies trusted data repositories. Its objective is to "promote universal and equitable access to quality-assured scientific data, data services, products and information, with a view towards long-term data stewardship", and to "fostering compliance with agreed-upon data standards and conventions, and providing mechanisms to facilitate and improve access to data". The Principles require that data be fully and openly shared, in accordance with international standards of ethical research conduct; made available with minimum time delay and free of charge; that all who produce, share, and use data work to preserve authenticity, quality, and integrity of the data, respect the data source and its privacy; that used data are appropriately cited and their originators acknowledged; and finally that data are labelled "sensitive" or "restricted" only with appropriate justification.

Open Data in a Big Data World is an international accord issued jointly by ISC (then ICSU and the International Social Science Council, ISSC), the InterAcademy Partnership (IAP), and The World Academy of Sciences (TWAS). The accord proposes 12 principles "to guide the practice and practitioners of open data, focused on the roles played by scientists, publishers, libraries and other stakeholders, and on technical requirements for open data. It also assesses the "boundaries of openness"." It takes on the emergence of 'big data' as a major opportunity for scientific discovery, while observing that 'open data' will "enhance the efficiency, productivity and creativity of the public research enterprise and counteract tendencies towards the privatisation of knowledge", that concurrent open publication of the data underpinning scientific papers can provide the basis of scientific 'self correction', and that maximising the benefits of big data "will depend on the extent to which there is open access to publicly-funded scientific data". Other concerns mentioned are to add to the stock of knowledge and understanding that are essential to human judgements, innovation and social and personal wellbeing; to enhance scientific productivity and creativity; and permit data and ideas to flow openly, rapidly and pervasively.

The 12 Principles (of which most are multi-faceted) describe the roles of scientists, universities and research institutes, publishers, funders, libraries, and others, and include:

- Make data openly available (scientists)

- Make data that provide evidence for published scientific claims concurrently and publicly available in an intelligently open[4] form (scientists)
- Require intelligently open access to the data concurrently with the publication which uses them, and require the full referencing and citation of these data (publishers)
- Regard the costs of open data processes as an intrinsic part of the cost of doing the research (funding agencies)
- Ensure that data are available to those who wish to use them and accessible over the long term (libraries, archives and repositories)
- Open data should be the default position for publicly funded science, with exceptions limited to issues of privacy, safety, security and to commercial use in the public interest
- Reused data should be cited with reference to their originator, to their provenance and to a permanent digital identifier
- Both data and metadata should be interoperable to the greatest degree possible
- If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be re-used with no more arduous requirement than that of acknowledging the producer.
- Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value

World Meteorological Organisation

The World Meteorological Organisation (WMO), which is an agency of the United Nations, was among the first global organisations to acknowledge the need for free and unrestricted exchange of data. Its current data policies are anchored in Resolution 40, which was approved by the WMO Congress in 1995. Resolution 40 reaffirms the commitment to free and unrestricted international exchange of meteorological data and also notes the increasing requirement for the global exchange of all types of environmental data. Among other rationales WMO emphasises the fundamental importance of unrestricted data exchange for the provision of meteorological services and for the ability of its member organisations to provide universal services in support of safety, security and economic benefits.

Under Resolution 40, WMO adopts as a fundamental principle to commit itself and its member nations to broadening and enhancing the free and unrestricted international exchange of meteorological and related data and products. WMO also obliges its members to provide free and unrestricted access to all data and products exchanged under the auspices of WMO to the research and education communities, for their non-commercial activities.

[4] "Intelligently open data" is a concept presented in the Royal Society report "Science as an open enterprise" (2012), implying that data should be accessible, intelligible, assessable, and usable by others. The concept has largely been superseded by the more recent FAIR principles.

The WMO data policies are currently (2020) being updated, not to change the fundamental principles of Resolution 40, but to cover new domains and developments.

Intergovernmental Oceanographic Commission

Another UN body with a similar role to the WMO is The International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission of UNESCO (IOC), which was established in 1961. Its purpose is to enhance marine research, exploitation and development, by facilitating the exchange of oceanographic data and information between participating member states, and by meeting the needs of users for data and information products. There are now over 80 oceanographic data centres working together to meet the IODE objectives which are centred around: facilitating and promoting the discovery, exchange of, and access to, marine data and information; encouraging the long term archival, preservation, documentation, management and services of all marine data, data products, and information; and developing or using existing best practices for the discovery, management, exchange of, and access to marine data and information.

The IOC Oceanographic Data Exchange Policy (revised 2019) is anchored in WMO Resolution 40, and obliges the IOC member states to "provide timely, free and unrestricted access to all data, associated metadata and products generated under the auspices of IOC programmes", and encourages the same for other data that are "essential for application to the preservation of life, beneficial public use and protection of the ocean environment, the forecasting of weather, the operational forecasting of the marine environment, the monitoring and modelling of climate and sustainable development in the marine environment".

OECD

The Organisation for Economic Co-operation and Development (OECD) has developed the OECD Principles and Guidelines for Access to Research Data from Public Funding, which were made an OECD Recommendation and endorsed by the OECD Council in 2006 and thus considered international "soft law". Since their publication, the OECD principles have been of particular influence on research funders across countries and research disciplines[5].

The underlying and stated intention of this OECD document is to increase the return on public investments in scientific research. A series of consequent societal benefits are identified:

- Good stewardship of the public investment in factual information;
- Creation of strong value chains of innovation;
- Enhancement of value from international co-operation.
- Reinforce open scientific inquiry;
- Encourage diversity of analysis and opinion;
- Promote new research;
- Make possible the testing of new or alternative hypotheses and methods of analysis;

[5] Current Best Practice for Research Data Management Policies, CODATA 2014

- Support studies on data collection methods and measurement;
- Facilitate the education of new researchers;
- Enable the exploration of topics not envisioned by the initial investigators;
- Permit the creation of new data sets when data from multiple sources are combined.

The OECD Principles are designed to promote data access and sharing among researchers, research institutions, and national research agencies, while recognising diverse national laws, research policies and organisational structures of its member countries. A set of 13 principles are laid out: Openness, Flexibility, Transparency, Legal conformity, Protection of intellectual property, Formal responsibility, Professionalism, Interoperability, Quality, Security, Efficiency, Accountability, and Sustainability - giving particular emphasis to openness as a goal.

Specifically, the OECD emphasises practices such as promoting a culture of openness and sharing of research data among public research communities; raising awareness about costs and benefits of restrictions and limitations on access to and the sharing of research data from public funding; and offering recommendations to member countries on how to improve the international research data sharing and distribution environment.

Group on Earth Observations (GEO)

GEO, the Group on Earth Observations, is a global, intergovernmental partnership working to improve access to and reuse of open earth observations through data sharing. A central part of its mission is to build the Global Earth Observation System of Systems (GEOSS), which includes a comprehensive data portal. In 2015, the GEO Principals endorsed a new set of Data Sharing Principles, which promote 'Open Data by Default'.

In their 2015 report on The Value of Open Data Sharing, ICSU CODATA and GEO presented a wide range of reasons for a transition from restricted to more open data policies for government data. The report highlighted several major trends "that have made the open and unrestricted uses of public data available through the GEOSS portal essential", and then proceeded to explore in some detail a range of benefits under five headlines:

- Broad economic benefits
- Enhancing social welfare
- Growing research and innovation opportunities
- Facilitating education
- Effective governance and policy making

More specifically, the data sharing principles state that "The societal benefits arising from Earth observations can only be fully achieved through the sharing of data, information, knowledge, products and services", and aims to "ensure that data and information of different origin and type are comparable and compatible, facilitating their integration into models and the development of applications to derive decision support tools".

The fundamental principle is that data, metadata and products will be shared as open data by default, subject to the conditions of user registration[6] and attribution when the data are reused. When sharing as open data is legally precluded, data should be made available with minimal restrictions on use and at no more than the cost of reproduction and distribution.

The GEO data sharing principles are expanded upon in the GEOSS data management principles, which include ten individual principles under the headlines Discoverability, Accessibility, Usability, Preservation, and Curation. The data management principles are further explained in a 40 page Data Management Principles Implementation Guidelines.

IPY Data Policy

The International Polar Years (IPY) are collaborative, international efforts of intensive research in the polar regions that have happened at 25-50 year intervals since 1882-1883. Given the long intervals, data legacy has been an important aspect of IPY, with corresponding emphasis put on data preservation and long-term accessibility. For the IPY 2007-2009, a specific IPY Data Policy was developed, in support of the overarching objectives of the IPY; to "ensure that data usability is a primary objective", and to "ensure the security, accessibility and free exchange of relevant data that both support current research and leave a lasting legacy".

The fundamental element of the IPY Data Policy was that all IPY data, including operational data delivered in real time, should be "made available fully, freely, openly, and on the shortest feasible time scale", with exceptions admitted only to protect the confidentiality where human subjects are involved, to protect the rights of the knowledge holders where local and traditional knowledge is concerned, and when data release might cause harm (e.g. to endangered species or sacred sites).

Further requirements of the IPY Data Policy were that IPY projects have an appropriately funded data management plan, provide complete metadata, ensure long-term preservation and sustained access, and acknowledge data authors.

The IPY Data Policy was also one of the first international data policy documents to lay down the principle of data acknowledgment: "*To recognize the valuable role of data providers (and scientists who collect or prepare data) and to facilitate repeatability of IPY experiments in keeping with the scientific method, users of IPY data must formally acknowledge data authors (contributors) and sources. Where possible, this acknowledgment should take the form of a formal citation, such as when citing a book or journal article.*"

The IPY Data Policy was a seminal document that later formed the shared basis for the SCAR Data Policy, the SOOS Data Policy, and the IASC Data Statement.

[6] User registration is stated as permissible for the GEOSS Data-CORE pool of datasets, but not encouraged.

Regional organisations

Antarctic Treaty (1961)

Countries working in the Antarctic operate within the framework of the Antarctic Treaty System. The cornerstone of the system is the Antarctic Treaty, which was signed December 1, 1959, and came into effect on June 23, 1961. Of particular relevance for polar data management and delivery is Article III, section 1(c), which stipulates that "scientific observations and results from Antarctica shall be exchanged and made freely available".

This Article has been followed up by ATCM Resolutions, such as:

- ATCM Recommendation XIII-5 (1985), which invites SCAR to offer advice "on steps that possibly could be taken to improve the comparability and accessibility of scientific data on Antarctica."
- ATCM XXII Resolution 4 (1998), which recommends that Consultative Parties establish National Antarctic Data Centres and link these to the Antarctic Data Directory, and that they give priority consideration as to how the requirement for freedom of access to scientific information is achieved within their national data management systems.

Agreement on Enhancing International Arctic Scientific Cooperation (2017)

No similar framework exists in the Arctic, as land-based and coastal research in the Arctic always happens within national jurisdictions. However, the Arctic Council member nations in 2017 signed an 'Agreement on Enhancing International Arctic Scientific Cooperation' for the purpose of increasing "effectiveness and efficiency in the development of scientific knowledge about the Arctic." Obligations that the parties agreed to include to:

1. Facilitate access to scientific information.
2. Support full and open access to scientific metadata; encourage open access to scientific data and data products and published results with minimum time delay, preferably online and free of charge or at no more than the cost of reproduction and delivery.
3. Adhere to commonly accepted standards, formats, protocols, and reporting.

Improved access to Arctic research and environmental monitoring data has been a recurring theme during the Arctic Science Ministerials held by the eight Arctic states and others[7].

European Union (2019)

The European Union has been introducing legislation, infrastructure, and other measures for open access to public data for more than two decades, with a notable milestone in the 'Public Sector Information Directive' in 2003. In 2019, this was replaced by the Open Data

[7] See Joint Statement of Ministers from the First Arctic Science Ministerial, Statement from the Second Arctic Science Ministerial

Directive, which is legally binding on its member states. Article 10 relates to research data and includes the following statements:

> *1. Member States shall support the availability of research data by adopting national policies and relevant actions aiming at making publicly funded research data openly available ('open access policies'), following the principle of 'open by default' and compatible with the FAIR principles. In that context, concerns relating to intellectual property rights, personal data protection and confidentiality, security and legitimate commercial interests, shall be taken into account in accordance with the principle of 'as open as possible, as closed as necessary'. Those open access policies shall be addressed to research performing organisations and research funding organisations.*
>
> *2. (...), research data shall be re-usable for commercial or non-commercial purposes in accordance with Chapters III and IV, insofar as they are publicly funded and researchers, research performing organisations or research funding organisations have already made them publicly available through an institutional or subject-based repository. In that context, legitimate commercial interests, knowledge transfer activities and pre-existing intellectual property rights shall be taken into account.*

In response to the transformative impact of digital technologies, the European Union has developed a "European strategy for data", aiming to promote a data-driven economy and innovation for citizen benefit. The strategy emphasises compliance with the EU's strict data protection rules.

Other developments and policy drivers

Data policies evolve in conjunction with the continual technological and institutional changes impacting the world of science and scientific data management. In the following sections we briefly present some recent developments that will, should, or may put new requirements on data policies.

The drive towards open data

Full and open access to research data is a common element of all the cited data policies. The open data principle is grounded both in public and societal benefits and in scientific justifications. The OECD Principles establish that publicly funded research data should be regarded as a public asset, and aim to maximise their benefit to society. Scientific justifications are tied to the need to promote scientific cooperation and scientific advancement, to improve the efficiency and quality of science, to induce proliferation of ideas, and to enhance the scientific productivity of data. The latter is of particular interest to polar research, where data collection tends to be prohibitively expensive. Another concern, perhaps more profound, is that open and concurrent access to all data supporting scientific claims is required for transparency and reproducibility in science. This is indeed touched upon by many of the mentioned data policies.

Although it has been widely recognised that open sharing of research data provides extensive benefits to science and society in general, the benefits for the investigator who makes his or her data available have been less obvious. However, as datasets are increasingly being published independently, there is growing recognition that published datasets constitute valuable scientific products in their own right[8]. There is also evidence that sharing detailed research data is associated with increased citation rates[9], implying that data exposure leads to increased scientific productivity.

The importance of continued open access to data has led to the assertion that the costs of open data and data management should be regarded as intrinsic parts of the cost of doing the research (Science International, IPY), and even that "it is a false dichotomy to argue that there is a choice to be made between funding provision for open data and funding more research. The practice of open data is a fundamental part of the process of doing science properly, and cannot be separated from it" (Science International, 2015).

Limits to openness and timeliness

At the same time it has been generally accepted that data cannot always be open. Most data policies recognise legitimate reasons for restricted access, which is reflected in wording like "as open as possible, as limited as necessary", or "ethically open". In a governance context, such reasons may relate to international relations; national security; law enforcement; legitimate commercial interests, such as trade secrets; and similar.

In a scientific context, the listed and valid reasons for restricted access will more typically include privacy and confidentiality when human subjects are involved, or in cases where data release may cause harm, e.g. by revealing locations of endangered species, cultural artefacts, or sacred sites. Restrictions may also be called for in protection of indigenous peoples' rights or to avoid compromising rights of the knowledge holders where local and traditional knowledge is concerned.

A separate question concerns the timing of data release. Some data policies allow researchers a certain period of privileged use of the data they have collected to enable them to publish the results of their research and to get appropriate recognition. The duration of privileged use varies and is a topic of debate, where rights of investigators must be balanced against concerns about restricting the scientific value of the data. It is argued that closing the data access prevents data reuse and thus their scientific productivity, creates inertia, limits scientific progress, and spoils opportunities for collaboration. There is no universal agreement on what is an appropriate delay between collecting the data and making the data open, and the policy limits seem generally to range from immediate release to two years. It should be kept in mind that several research communities have demonstrated substantial benefits of immediate data release (Science International, 2015).

[8] https://www.nature.com/articles/sdata2018259
[9] Piwowar & al. 2007, Piwowar & al. 2013

The 'data deluge' and 'big data'

Science, just as much as the world in general, is undergoing a 'digital revolution' where rapid growth in computing power and data storage capacity is shaping many aspects of both professional and daily lives. We are seeing an unprecedented explosion in the capacity to acquire, store, manipulate and near-instantaneously transmit vast and complex data volumes.

The "internet of things" permits independent devices on all scales to collect data from their environment, constantly opening new opportunities for research. Humans are leaving electronic traces wherever they go; traces that are being collected and turned into vast, complex datasets. Huge datasets can be subjected to big data analysis, allowing the detection of patterns that were undiscoverable without today's computing power. "Cloud computing" disconnects data from their physical origin and provides computing power independent of location. Big data analyses, through tiers of analyses and meta-analyses, are prone to obscure the provenance of the base data. "Linked data" allow separate datasets to be semantically linked in ways that permit a computer to identify deeper relationships between them, connecting related data that were not necessarily designed for mutual integration - as long as the data are openly available and free to be linked.

This digital revolution raises some data policy challenges as well as ethical concerns. Scientific datasets and data collections have generally been managed and published as discrete entities, with metadata, licences, and authorship assigned to the dataset as a whole rather than individual data points. This basis for attribution and provenance tracing will easily break down in a world of digitally networked and big data, thus creating a need for new ways to ensure traceability and transparency, and perhaps new ways to perceive data resources and acknowledge authorship. It has been observed that "the challenges associated with providing recognition to the generators of datasets integrated into complex data products, a phenomenon of data-intensive research, means that many authorities argue that licences such as CC-BY that require attribution are not sustainable or appropriate in a Big Data age." (Science International, 2015). The same source points out that "The veracity and the peer review of results based on big data, however, pose severe problems for effective scrutiny, with a clear need to establish a reproducibility standard."[10]

Another challenge is linked to the increasing ability of all internet users to produce and distribute exact - or non-exact - reproductions of digital material, including protected works. This is changing the intellectual property landscape and raising new challenges in tracing provenance and authenticity.

[10] Science International (2015): Open Data in a Big Data World. Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP)

Big data may even require personal data protection beyond conventional anonymisation because data on individual behaviour may emerge from pattern recognition, thus compromising the privacy of individuals.

Commercial and industrial partnerships

While much of the data created by private funders, including commercial operators, is subject to commercial considerations and does not have the same ethical requirements for data sharing that publicly funded data have, much public benefit can come from sharing this data with the public, where possible. Where data centres engage with commercial operators and other private data owners, vast pools of data may become available as a valuable resource for scientific research.

Data and results from publicly funded research may also form a basis for commercial enterprise and innovation. This is commonly encouraged by governments and funding agencies, but will in some cases require careful consideration of legal rights and licencing.

New cost models and big data infrastructure costs

Most of the data policies we have examined state that data should be "freely available", in some cases modified to "available at no more than the cost of reproduction and distribution". While the development of a modern digital infrastructure has largely annihilated the distribution costs for modest data volumes, the situation becomes different for 'big data' because of the extensive bandwidth requirements. Some commercial data repositories that are hosting research data may also have cost models where data storage is inexpensive while bandwidth usage incurs substantial costs[11].

ISC and other bodies argue that the costs of open data processes be regarded as an intrinsic part of the cost of doing the research (Science International, 2015), and thus funded as such. The principle clearly applies to all fixed and ordinary costs associated with data management. However, for big data there may be usage-dependent costs that cannot reasonably be funded as part of the original research grant or the operating budget of the data centre. In some such cases bandwidth costs may be reduced or eliminated by allowing users to process the data where they reside instead of moving the data ("bring the algorithms to the data"). To the greatest extent possible, data should be made available without cost, save for exceptional circumstances where network charges or other significant costs cannot be reasonably borne by the data provider.

[11] The EU Open Data Directive also allows for costs connected with anonymisation of personal data and measures taken to protect commercially confidential information.

The FAIR Principles

The FAIR Principles were first presented in an important publication (Wilkinson et al., 2016) that has significantly influenced data sharing and data policy developments. The paper was motivated by a need to define 'good data management' in a sense that would facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration, and analysis of task-appropriate scientific data and associated algorithms and workflow. The FAIR principles put specific emphasis on enhancing the ability of machines to automatically find and use data.

The FAIR principles assert that data collections should be Findable, Accessible, Interoperable, and Reusable, and each of the four are translated into specific requirements to the data management system. Findability refers to the capacity to search for and discover data collections, and involves requirements on metadata, identifiers, and indexing. Accessibility is a measure of the ease with which information can be directly obtained or accessed once discovered. Interoperability is the degree to which independent data sets can be combined and integrated with one another, which can be facilitated by using consistent standards, encoding, and vocabularies. Reusability means that the data can be put to multiple uses beyond its original purpose, and includes requirements on usage licences, provenance, and community standards.

The principles refer to three types of entities: data, metadata, and infrastructure. While the FAIR principles have made their way into a large number of data policies, it is important to remember that full implementation across all three entities will incur significant costs. Full adherence to the FAIR principles for the 'long tail of research data'[12] may not even be possible. However, the principles also represent best practices for data management and can be implemented along a continuum from unstructured, undocumented data to fully FAIR data. The balance between the utility of fully FAIR data and the cost of implementing it must be kept in mind when introducing the principles as a matter of policy.

A side benefit of FAIR data is that datasets released from individual data centres increasingly can be fed into federated data sharing networks, allowing for aggregation, subsetting, and searching, regardless of the origin of a particular dataset. This opens for more flexible and capable dataset search systems than the traditional, monolithic data catalogues.

The TRUST Principles

The TRUST principles have emerged as a Research Data Alliance community effort and were published in the 2020 article "The TRUST Principles for digital repositories" (Lin et al. 2020). The principles apply to digital data repositories and are intended to ascertain their trustworthiness, especially for those responsible for the stewardship of research data.

[12] I.e. the vast amount of small, non-standardised, and often poorly documented datasets from small-scale projects. See [PDF] Shedding Light on the Dark Data in the Long Tail of Science.

The acronym signifies Transparency (about specific repository services and data holdings that are verifiable by publicly accessible evidence), Responsibility (for ensuring the authenticity and integrity of data holdings and reliable, persistent services), User Focus (ensuring that the data management norms and expectations of target user communities are met), Sustainability (of services and data holdings, long-term), Technology (infrastructure and capabilities to support secure, persistent, and reliable services).

Indigenous Knowledge and Data Use and Stewardship Principles

Indigenous Knowledge is a systematic way of thinking and knowing that is elaborated and applied to phenomena across biological, physical, cultural and linguistic systems. Traditional Knowledge is owned by the holders of that knowledge, often collectively, and is uniquely expressed and transmitted through indigenous languages. It is a body of knowledge generated through cultural practices, lived experiences including extensive and multigenerational observations, lessons and skills. It has been developed and verified over millennia and is still developing in a living process, including knowledge acquired today and in the future, and it is passed on from generation to generation (Indigenous Peoples Secretariat, 2015).

Indigenous Peoples' *data* include data generated by Indigenous Peoples, as well as by governments and other institutions, on and about Indigenous Peoples and territories. This includes information about Indigenous communities and the individuals, Indigenous and non-Indigenous, that live within. Indigenous peoples and their representative organization have established principles for appropriate and ethical use of Indigenous data[13]. At a national scale, such principles have been developed or are under development, including the Canadain First Nation's Ownership Control Access and Possession OCAP (Schnarch, 2004) and emerging principles being established under the Canadian National Inuit Strategy on Research (ITK 2018). Particularly notable are the CARE Principles for Indigenous Data Governance that were drafted in 2018 and introduced by the Global Indigenous Data Alliance in 2019 on the premise that the movement toward open data and open science does not fully engage with Indigenous Peoples rights and interests (Carroll et al 2020)[14]. The principles pertain to the management of data about and collected by Indigenous people, and stipulate Collective benefit from the data, Authority to control such data, Responsibility to support, and Ethical processes.

Indigenous data principles and practices are evolving rapidly and warrant particular attention by the polar data community. An additional effort complementary to this paper and led by members of the Arctic data community will focus on enhancement of research

[13] Carroll, SR, Rodriguez-Lonebear, D and Martinez, A. 2019. Indigenous Data Governance: Strategies from United States Native Nations. Data Science Journal, 18(31): 1–15. DOI: https://doi.org/10.5334/dsj-2019-031

[14] Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R. and Sara, R., 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal, 19(1).

data practices based on established Indigenous principles. This effort will be further developed under the SAON ROADS Process (2020)

Demand for transparency

Data transparency corresponds with the scientific principles of repeatability and reproducibility. For several reasons, including a few notorious cases of falsified, high-profile scientific results[15], scientific journals are increasingly requiring that all data supporting a scientific work are cited and made openly available[16]. Examples can be found at, i.e., Springer Nature and sciencemag.org. Correspondingly, some of the global data policy statements demand that data providing evidence for a scientific claim must be published concurrently and publicly available. The recommendation from ISC is that such data should be published in a way that "permits the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations". (Science International 2015).

Journal publishers and editors have also been realising that providing direct access to the data increases the appeal of the journal. However, the requirement for concurrent publication of articles and supporting data has led to examples of sub-optimal data publication practices. If data are accessible only as poorly described 'supplementary materials' in unsuitable formats, or as limited subsets of the original datasets, they will not be reusable as desired. The practice may even impede proper dataset publication.

Other ethical considerations

Different communities of practice may have different data-related norms, protocols or policies. For example, some disciplines within the social sciences may have very specific protocols required by formal research ethics processes and/or the nature of their research and the ethical dimensions that they must consider (e.g. the IASSA Research Principles, the NSF Arctic Horizons report, the OECD policy paper Research Ethics and New Forms of Data for Social and Economic Research, and the RDA Ethics and Social Aspects of Data IG.

Responsible reuse of data requires that users become familiar with the specific context of data production, access and reuse to avoid misusing data. This includes fully open and the more restricted forms of data discussed. Describing the implementation of ethical practices is beyond the scope of this paper. The authors are working with the broader polar data community to further develop shared practices through processes such as SAON ROADS (2020) process. Readers are also directed to cited publications (e.g.King 2011[17], Indigenous Data Sovereignty, Pulsifer et al. 2011[18]).

[15] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2685008/
[16] http://www.ijdc.net/article/view/12.1.65/467
[17] King, G. (2011). Ensuring the data-rich future of the social sciences. science, 331(6018), 719-721.
[18] Pulsifer, P. L., Laidler, G. J., Taylor, D. F., & Hayes, A. (2011). Towards an Indigenist data management program: Reflections on experiences developing an atlas of sea ice knowledge and use. The Canadian Geographer/Le Géographe canadien, 55(1), 108-124.

Better legal instruments for data sharing

Over the last two decades, the research community has gained access to new legal instruments suitable to open data sharing, with the emergence of several licences that have gained worldwide recognition. The Creative Commons licences, in particular the attribution licence (CC-BY), are notable examples. Creative Commons was established in 2001, and the number of CC-licenced works started to grow considerably after 2010[19]. Open attribution licences have made it easier to share scientific data openly and gain recognition for data as contributions to the scholarly record. However, copyright legislation and specific requirements and obligations tied to licencing vary across jurisdictions. Thus, international data policies must have some flexibility in their licencing requirements.

Metrics

With data citations becoming common practice in scholarly publishing, datasets are also becoming regarded as valuable science products in their own right. This opens discussions about ranking of datasets by scientific productivity or impact. Counting the number of downloads is a traditional but crude measure, and dataset citations are gradually becoming a more common factor by which data are assessed as research contributions. Citations are also becoming an incentive for data sharing, although dataset citations usually do not carry the same weight as citations of scientific papers. Funding agencies are, however, starting to explore the scientific productivity of datasets as an element to factor into funding considerations, as a way to promote publication and early release of research data.

Various other ways to measure the scientific impact of datasets have been suggested. What they seem to have in common, is their reliance on linkages between datasets and the scientific results they generate. Persistent and unique identifiers are key to linking research data with scientific results and tracking data reuse, and so is a general dataset citation requirement.

Even with widespread adoption of persistent, unique identifiers and good citation practices, it is becoming increasingly difficult to equitably recognise the contributions of all dataset producers, due to the complexity and size of data sharing pathways. Traditional data citations work well for research papers that rely on datasets collected by one or a few scientists but tend to break down when working with large aggregate datasets that may include contributions from hundreds or thousands of researchers and projects. When citing the aggregate dataset, rather than add hundreds of entries to their reference list, the attribution from the original dataset creator is severed, possibly violating a CC-BY or similar licence, and the data provenance may be obscured. The data community must find sustainable and practicable solutions to allow appropriate attribution across all levels of data granularity, and also the tracking of impact of data publication or to amend expectations and licence conditions to support the use of derived and aggregate datasets. Systems to allow

[19] Creative Commons, State of the Commons

lists of DOIs to be appended to a research paper that in turn allow databases to track usage of individual data records is one potential future a solution to this problem but such infrastructure has not yet been established and widely adopted[20].

Summary: common principles

The principles and core requirements that already are or are becoming generally recognised by the above institutions should form the basis for alignment of the polar data policies. We will summarise those requirements and principles here. In the following, the terms “data policies” and “policy documents” refer to the policies that are identified in this document.

What emerges from all the policy documents is that full and open access to research data has become firmly established as an international norm for publicly funded research data. In the case of Antarctic research it is also a legal requirement following the Antarctic Treaty. The extent of the principle may differ, but a common wording is "open by default and design” (ISC, GEO, EU). The principle is often combined with assertions that data access should be “free”, “timely”, and “unrestricted”, or that data can be reused with "no more arduous requirement than that of acknowledging the producer", or at no more than the cost of reproduction and distribution.

Open data is the default position, but often with certain caveats which has led to wording like “as open as possible, as closed as necessary” or “ethically open”. It is generally recognised that restricted access can be justified for reasons of privacy, safety, security, environment protection, and ethical considerations, including protection of the rights of indigenous peoples. However, it is emphasised that data should not be labelled as sensitive or restricted without proper justification.

Many data policies have similar requirements for interoperability, compatibility, adherence to standards, persistent and unique identifiers, and documentation (metadata and provenance). Such requirements are largely captured by the FAIR principles, which can be regarded as the most updated set of usability requirements. A higher ambition is to enable cross-linkages of datasets, originators, publications, and other scientific artefacts. When open data can be linked with other data based on context and content it will maximise their semantic value.

Common policy elements that are not captured by the FAIR principles are data curation and management, long-term preservation and sustained access. Data preservation requirements include preservation of integrity, quality, and authenticity, and the ability to trace provenance and authenticity. A more specific requirement is that data should be archived in their most usable form. This at least partly relates to the practice introduced by some

[20] For a deeper investigation of these issues we will refer to Task Group on Data Citation Standards and Practices, C.-I., 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, pp.CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043

scientific journals, where data supporting scientific papers are released as poorly described “supplementary materials” that often contain only limited subsets of larger and more useful datasets.

Most data policies state as principles that data originators should be acknowledged for their effort and that datasets should be (formally) cited when reused. This is becoming common practice, and relies on an underlying infrastructure of, inter alia, sustainable data repositories, persistent and unique identifiers, cross-linkages, provenance documentation, and perhaps - in the ‘big data’ scenario - new reproducibility standards. Such infrastructure also serves the interest of transparency and traceability.

Licencing is generally not mentioned in the global data policies, in many cases because the policies predate the widespread application of data licences. However, unlicenced data, even if it is open data may, be rendered unusable if no licence is attached. In some jurisdictions no licence is regarded as the same as 'all rights reserved', thus restricting any reuse to very limited circumstances. Licencing thus seems like a necessary policy element, and is indeed a part of the FAIR principles.

Part 3, Core Principles

This data policy paper was prepared with reference to the data policies listed in table XXX and attempts to be compatible with them all. Where there is conflict between two or more relevant data policies, the data producer or user should use their discretion in choosing the most ethical and practicable path.

Document	Latest revision date
Antarctic Treaty	1959
Antarctic Treaty Resolution 4	1998
World Meteorological Organisation Policy and Practice for the Exchange of Meteorological and Related Data and Products Including Guidelines on Relationships in Commercial Meteorological Activities	1999
Scientific Data and Information Report of the CSPR Assessment Panel	2004
International Polar Year 2007-2008 Data Policy	2006
Organisation for Economic Cooperation and Development Principles and Guidelines for Access to Research Data from Public Funding	2006
Scientific Committee on Antarctic Research Data Policy	2011
Statement of Principles and Practices for Arctic Data Management	2013
Southern Ocean Observing System Data Policy	2015

Open Data in a Big Data World: An International Accord	2015
World Data System Data Sharing Principles	2015
Global Earth Observation System of Systems Data Sharing Principles	2015
The FAIR Guiding Principles for scientific data management and stewardship	2016
Agreement on Enhancing International Arctic Scientific Cooperation	2017
CARE Principles for Indigenous Data Governance	2018
Intergovernmental Oceanographic Commission Oceanographic Data Exchange Policy: Status Report	2019
American Geophysical Union Position Statement on Data	2019
The Beijing Declaration on Research Data	2019
European Union Open Data Directive	2019
European Strategy for data	2020
Proposal for a Regulation on European data governance (Data Governance Act)	2020

Key objectives

That publicly funded research data, including nearly all research data from the polar regions, should be regarded as a public asset and managed in a way that will maximise their benefit to society has become an almost universal presumption of global scientific organisations, governments and funding agencies. Like the IPY data policy, updated polar data policies should aim to "provide a framework for these data to be handled in a consistent manner, and to strike a balance between the rights of investigators, the rights of indigenous peoples, and the need for widespread access through the free and unrestricted sharing and exchange of both data and metadata."

More specific objectives for research data policies will be to promote scientific cooperation and scientific advancement, to improve the efficiency and quality of science, and to enhance the scientific productivity of data. The latter is of particular interest to polar research, where data collection tends to be prohibitively expensive and duplication of data collection efforts are correspondingly undesirable. Even more importantly, the management of polar data and all other research data should serve to ensure transparency and reproducibility in science, and to preserve scientific legacies over the long-term.

Improving the scientific productivity of research data allows users to generate more knowledge per collected dataset, but it requires a more streamlined data flow. Good data

policies can promote this by stipulating best practice requirements wherever it can be observed that current practices are impeding the free and open exchange of research data. By the same measures we can hope to highlight gaps in knowledge, induce innovation and the proliferation of ideas, and stimulate the search for new knowledge.

Recommended core principles for all polar data policies

Scientific advancement depends on cooperation among researchers, policy makers, government, rights holders, residents, and other members of the public, crossing scientific disciplines and national boundaries. International data policies should serve to facilitate such collaboration. The following sections present a set of fundamental principles that are widely acknowledged in global and regional data policies, which we believe should form the core of polar data policies as well. This set of agreed principles is aimed to provide a foundation for an aligned view of how polar data and information should be curated, managed, and delivered. We have worded the principles in a way that should be suitable for direct inclusion in formal, polar data policy documents, with only minor modifications dependent on local context (such as the exclusion of the reference to the Antarctic Treaty in Arctic documents).

Members of the Arctic, Antarctic, and Southern Ocean science communities work in nations, institutions, and disciplines that have varied laws and research policies. Data centres, funding agencies, and research institutions are encouraged to develop more specific policies and procedures to implement the policy elements contained in this document in a manner that aligns appropriately with more local policy and legal requirements.

<Data must be ethically open>

Data from publicly funded research should be open by design and by default in order to release their full potential as a primary resource for knowledge discovery. Full, free, and open access for all users should be the norm unless there are valid reasons for restricted access. For Antarctic research data, this is also a requirement of the Antarctic Treaty. This principle may be referred to as “as open as possible, as closed as necessary” or as ethically open data.

ICSU (2004) defines “Full and open access” as equitable, non-discriminatory access to all data. Open data as a concept is generally understood to denote data in an open, platform-independent format that can be freely used, re-used and shared by anyone for any purpose.

It is generally recognised that sharing and use of some data must remain partially or completely limited for ethical, cultural or legal reasons (IPY 2006, IASC 2013, CARE 2019). Valid reasons for such limitations may relate to privacy where human subjects are involved, safety, security, environment protection, and other ethical considerations, including

protection of the rights of indigenous peoples. However, it is emphasised that data should not be labelled as sensitive or restricted without proper justification.

<Data should be free>

The distribution and reuse of research data should be free of charge, and delivered at no more than the cost of reproduction and delivery. With modern digital communication technologies the distribution costs for modest data volumes have largely been eliminated, and typically do not justify any cost recovery on the distributor side. The costs of open data processes should be regarded as an intrinsic part of the cost of doing the research, and thus funded as such.

However, the handling of large data volumes ('big data') may incur significant costs, primarily due to bandwidth requirements. Where such usage-dependent costs cannot reasonably be funded as part of the original research activity or the operating budget of the data centre, or avoided by performing the data analyses without moving the data, some cost recovery may be justified even under a free and open access data policy.

<Data must be provided in a timely manner>

To facilitate reuse of data while they are most valuable, all research data should be made available as soon as possible after their collection and if possible near real-time. Some latency may be required for data processing, quality control, compilation of well-documented and FAIR data products, and, in some disciplines, formal peer review of initial scientific findings.

Some data policies allow researchers a certain period of privileged use to facilitate publication and recognition, through an embargo on data publication. Such data embargoes should be applied only for good cause and for the shortest time feasible to allow for good data processing practices and to respect the scientific endeavours of data creators. When embargoes are considered, it is important to evaluate the broader benefits of immediate release, and to consider the negative effects of embargoes on scientific productivity. A maximum embargo limit should be stipulated, and embargoed data should include a date for review of their embargoed status, along with documented reasons for the embargoed status.

<FAIR Principles should be applied to the greatest extent practicable>

To ensure the efficient and effective uptake of data, the FAIR principles (Wilkinson et al. 2016) must be followed to the greatest extent practicable and ethical ("FAIR as far as possible"). The FAIR principles assert that data collections should be Findable, Accessible, Interoperable, and Reusable. These principles depend on community-agreed formats, languages, and vocabularies for both data and metadata.

The FAIR principles involve technical requirements that may be costly to implement. For this reason it is unrealistic to make all data fully FAIR, especially if we consider “the long tail” of research data.

When unrestricted open access is unethical or otherwise inappropriate the FAIR principles envisage creation of different user roles and mechanisms for user verification to provide controlled access.

The FAIR principles put specific emphasis on enhancing the ability of machines to automatically find and use the data. However, the principles also represent best practices for data management and can be implemented along a continuum from unstructured, undocumented data to fully FAIR data. Findability and online data accessibility should be regarded as universal requirements. Some FAIR elements are important also when considering data reusability in general, and will be reiterated in the following as universally important requirements for the long-term management of research data.

<All data must be accompanied by a complete set of metadata>

Structured, standardised metadata are essential to the discovery, access, and effective reuse of data, allowing users to assess the quality of the data and any processing that has been applied to it. All data must be accompanied by a full set of metadata that appropriately documents and describes the data. Metadata elements should provide a clear description of the data; their provenance, the data structure; calibrations; and methods, including units, associated errors, or other limitations where possible. Shareable metadata, with sensitive details obscured or generalised, must always be available, even when the data themselves cannot be made publicly available for ethical or practical reasons.

More specific metadata requirements are included in the FAIR principles.

<Data should have persistent and globally unique identifiers>

Persistent and globally unique identifiers (PIDs) should be used for all data and remain linked to the data through republication or data aggregation processes. Unequivocal dataset identification is key to long-term data preservation, identification, attribution, data citation, provenance tracking, linking research data with scientific results, and tracking of the distribution and impact of data collections. For data and research products this includes the use of Digital Object Identifiers (DOIs) and other persistent identifiers that can be applied to both datasets and observations. Other types of PIDs should be considered when helpful in managing the data, such as ORCIDs for researchers.

Further guidance on persistent identifiers is included in the FAIR principles.

<Data must be labelled as reusable>

Open data access and legal interoperability requires that the rights to reuse the data are made clear to the user. For this reason, the rights and obligations of the data originator and the data user should be declared by attaching a rights waiver, a public domain statement, or an internationally recognised data licence to the dataset. This should be a non-restrictive licence specifying that the data may be re-used and specifying no requirement more onerous than an acknowledgement of the data's source, e.g. the Creative Commons open attribution licence (CC-BY). Where possible, the rights waiver or licence should be assigned by the owner or source of the data, and these parties should be identified in accompanying metadata. Failure to label the data as reusable may render the data legally unusable in some jurisdictions.

Further metadata requirements are included in the FAIR principles.

<Data sources should be attributable and attributed>

Data citation is an essential element of good research practice. To recognise the valuable contributions of data providers and to enhance repeatability and transparency of research results, data users must formally acknowledge data authors and sources. In some cases, aggregated datasets may comprise contributions from large numbers of data producers. Data managers should investigate and develop best practice methods for citing such datasets. Where possible, authors should use and cite original data, not subsets or derivatives, to prevent fragmentation of attribution. Best practices for data citation are outlined in the Joint Declaration of Data Citation Principles (JDDCP)[21].

For data to be easily attributable they must have a persistent and unique identifier. The information attached to the citation and the identifier must allow the provenance of the data to be assessed. Data should be referenced by means of a citation including a permanent digital identifier, and should be curated in and accessible from a trusted repository.

<Data must be appropriately preserved for the long term>

Given that the long-term value of data may not be recognised until long after collection, preservation of data to ensure a lasting legacy of research programmes and projects is essential.

Data must be preserved in such a manner that it is resilient to corruption or loss. This requires ensuring that adequate backup procedures are in place, that metadata records are maintained, and that files and formats remain readable and free from damage and degradation over time. Data must be protected against unintentional and unauthorised

[21] See also A data citation roadmap for scientific publishers.

modifications. The use of open and well-documented formats is strongly encouraged to ensure that data are in a suitable form for long-term curation.

<Data management and long-term curation must be planned and resourced>

Proper planning of data management and long-term curation is an integral part of any scientific endeavour. Projects should develop data management plans in advance of collecting data that outline how any data captured, modelled or acquired will be managed both during the life of the program and beyond. Where possible, data should be deposited for long-term management in repositories that adhere to the TRUST principles.

Funding agencies and science managers must consider the long-term resource required to host and manage data beyond the project lifespan. This will involve consideration of hardware and software costs and the need for staff with specialist skills in data preservation, data curation, providing access to data and increasing interoperability between datasets.