



Earth Science Data Repositories: Implementing the CARE Principles

RESEARCH PAPER

MARGARET O'BRIEN 

RUTH DUERR 

RILEY TAITINGFONG 

ANDREW MARTINEZ 

LOURDES VERA 

LYDIA L. JENNINGS 

ROBERT R. DOWNS 

ERIN ANTognoli 

TALYA TEN BRINK 

NICOLE B. HALMAI 

DOMINIQUE DAVID-CHAVEZ 

STEPHANIE RUSSO CARROLL 

MAUI HUDSON 

PIER LUIGI BUTTIGIEG 

*Author affiliations can be found in the back matter of this article

 ubiquity press

ABSTRACT

Datasets carry cultural and political context at all parts of the data life cycle. Historically, Earth science data repositories have taken their guidance and policies as a combination of mandates from their funding agencies and the needs of their user communities, typically universities, agencies, and researchers. Consequently, repository practices have rarely taken into consideration the needs of other communities such as the Indigenous Peoples on whose lands data are often acquired.

In recent years, a number of global efforts have worked to improve the conduct of research as well as data policy and practices by the repositories that hold and disseminate it. One of these established the CARE Principles for Indigenous Data Governance (Carroll et al. 2020), representing ‘Collective Benefit’, ‘Authority to Control’, ‘Responsibility’, and ‘Ethics’ hosted by the Global Indigenous Data Alliance (GIDA 2023a).

In order to align to the CARE Principles, repositories may need to update their policies, architecture, service offerings, and their collaboration models. The question is how? Operationalizing principles into active repositories is generally a fraught process. This paper captures perspectives and recommendations from many of the repositories that are members of the Earth Science Information Partners (ESIPFed, n.d.) in conjunction with members of the Collaboratory for Indigenous Data Governance (Collaboratory for Indigenous Data Governance n.d.) and GIDA, defines and prioritizes the set of activities Earth and Environmental repositories can take to better adhere to CARE Principles in the hopes that this will help implementation in repositories globally.

CORRESPONDING AUTHOR:

Margaret O'Brien

University of California, Santa Barbara, CA, USA

margaret.obrien@ucsb.edu

KEYWORDS:

Data Repositories; CARE Principles; Indigenous Peoples; Collaboratory for Indigenous Data Governance; Earth Science Information Partners; ESIP

TO CITE THIS ARTICLE:

O'Brien, M, Duerr, R, Taitingfong, R, Martinez, A, Vera, L, Jennings, L L, Downs, R R, Antognoli, E, ten Brink, T, Halmi, N B, Carroll, S R, David-Chavez, D, Hudson, M, and Buttigieg, P L 2024 Earth Science Data Repositories: Implementing the CARE Principles. *Data Science Journal*, 23: 37, pp. 1–29. DOI: <https://doi.org/10.5334/dsj-2024-037>

The technological revolution of the 21st century has radically transformed our abilities to access, produce and transform data about virtually every aspect of our world. All data are influenced by their cultural and political context (and, thus, carry bias), an especially important concern for data related to Indigenous Peoples. Following Younging (2018) we use the terms 'Indigenous Peoples' to represent political and cultural collectives as used by the United Nations and 'Indigenous communities' to describe the breadth of 'nuanced, overlapping, and complex intra-group systems and sub-divisions' by which Indigenous Peoples organize themselves (Cultural Survival and First Peoples Worldwide 2023). We defer to Carroll et al. (2020) and the Royal Society Te Aparangi (2023) in treating 'Indigenous data' as information and knowledge, born digital or not, that relates to Indigenous Peoples and communities at the collective and individual scales. Obvious examples are data such as health records and socioeconomic research about Indigenous individuals. However, also included are data derived from Earth and environmental observations or samples, specimens, and museum collections, especially when these relate to Indigenous Peoples. This definition is by necessity, very broad. Recognizing that these data hold relevance to and may impact Indigenous communities is especially important since Indigenous communities have been disproportionately harmed and marginalized through research activities and associated materials, including data (Arbour and Cook 2006; Drabiak-Syed 2010; Hernandez 2022; MacInnes Colchester and Whitmore 2017).

The CARE Principles for Indigenous Data Governance were produced by the Research Data Alliance International Indigenous Data Sovereignty Interest Group and are 'grounded in community values, which extend to society at large' (International Indigenous Data Sovereignty IG 2017). They provide a mechanism to move towards equitable futures in Indigenous data governance and research, and complement the FAIR Data Principles (Wilkinson et al. 2016), by adding 'people- and purpose-oriented' goals (C, Collective benefit; A, Authority to control; R, Responsibility; and E, Ethics). CARE articulates high-level guidance to strengthen Indigenous data governance across research, government, and institutional settings by bolstering 'Indigenous control for improved discovery, access, use, reuse, and attribution in contemporary data landscapes' (Carroll et al. 2020). Although principles like the CARE suite might be relevant for various community contexts, here we focus on their original intent for Indigenous data, and promote their further development by offering concrete recommendations for environmental data management and stewardship. In practice, CARE Principles have been shown to improve community awareness and ethical standards in environmental disciplines around Indigenous data (Jennings et al. 2023).

The United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP, (United Nations, n.d.-b)) affirmed, among other rights, the right of Indigenous Peoples to self-govern. On this basis, movements for Indigenous data sovereignty (IDSov) promote the rights of Indigenous Peoples in the control of data about their peoples, lands, and resources, from collection to use and reuse. Concurrently, Indigenous data governance (IDGov) refers to the mechanisms that increase Indigenous control over data (Kukutai and Taylor 2016). IDSov, IDGov, and the CARE Principles are widely recognized as key to data practices that balance calls for open sharing of data with responsibilities to the Indigenous Peoples to whom those data relate.

The CARE Principles inform the work of the Earth Science Information Partners (ESIP), a group of data managers, practitioners and scholars whose vision is 'a world where data-driven solutions are a reality for all by making Earth Science data actionable by all who need them anytime, anywhere' (Shingledecker et al. 2024). To respond to the CARE Principles as part of our standards for storing and retrieving data, ESIP relies on the work laid out by Indigenous activists and scholars from Indigenous Data Sovereignty Networks (IDSNs) spanning the USA, Canada, Aotearoa (New Zealand), Spain, and Australia under the Global Indigenous Data Alliance (2023) and the Research Data Alliance interest group (International Indigenous Data Sovereignty IG 2017).

The guidelines presented here are written for data repository managers and data system architects, particularly those responsible for Earth and environmental science data. In general, such repositories steward data to make it available for the long-term, i.e., on timescales much longer than the working life of an individual researcher. These data repositories can include data from many scientific disciplines. Much of this data was collected in or are about

places, communities, or phenomena that are governed by or have relevance to Indigenous Peoples and cultures. Although following the CARE Principles would typically require specific acknowledgement and access, repositories may be entirely unaware of those connections.

Our aim is to provide repositories with guidelines to begin a discussion about implementing practices that promote proper handling of Earth and environmental data related to Indigenous Peoples. We provide background outlining the role of repositories in the data lifecycle and a brief description of the CARE Principles. The results are a high-level set of recommended practices that are aligned with the personnel roles typically represented in repository staff. They are not meant to be deeply comprehensive and so do not contain explicit solutions to every data management problem or data type a repository may encounter. These guidelines also do not cover research activities that occur prior to a repository becoming involved with a project, which often occurs near the project's end.

BACKGROUND

EARTH AND ENVIRONMENTAL DATA REPOSITORIES

In general, data repositories archive, curate and preserve information for the long term and make it available for use by a variety of audiences ([CCSDS 650.0-B-1 2012](#)). The kinds of data a repository handles can vary widely depending on the repository's mission and the needs of its users. While some repositories – such as Figshare and Zenodo – are generalist, others are limited to data from specific missions or projects sponsored by their funding agency, certain topics, scientific domains or regions, or the data from specific institutions. A repository may fall under several of these categories, e.g., the holdings of a domain-specific institutional repository may be limited both by its scientific scope and to researchers associated with that institution. Staffing varies widely in repositories as well – from small groups of a few, perhaps part-time employees, to large repositories with perhaps 100+ FTE ([Baker and Duerr 2017](#)).

Repository operations vary widely, and funding sources and contractual obligations determine many aspects of their operation, such as the level of curation provided for the data. In most cases, the content, quality and completeness of a dataset's metadata depend strongly on the researchers or agencies that provided it ([Baker and Duerr 2017](#)), and consequently, the adage 'garbage in – garbage out' often applies. Repositories' audiences and services provided also vary widely and evolve over time as recognition of the utility of their data changes. This often leads to a general broadening of the audiences served and more recently, increased recognition of the rights and needs of communities and individuals from which the data originate ([Baker, Duerr, and Parsons 2015](#)). This may result in plans for expanded services if support can be found.

Several organizations provide repository certification that cover the basic activities and responsibilities of data repositories ([Downs 2021](#)), in line with the TRUST Principles for Digital Repositories that recommend applying Transparency, Responsibility, User focus, Sustainability, and Technology to manage data ([CoreTrustSeal Standards and Certification Board 2022](#); [Lin et al. 2020](#)). Many Earth and Environmental repositories have or are in the process of undergoing certification by one or more of these organizations. However, such certification processes omit repository responsibilities that address the CARE Principles, although some of the recommendations here are related to these certification processes. Readers should already be familiar with the basic requirements of certification boards, as these often can form a basis to support specific activities related to CARE Principles.

Every repository has one or more designated communities that it was designed to engage both in terms of data acceptance and access support ([CCSDS 2012](#)). For the repositories represented in ESIP, historically this has primarily been environmental or Earth scientists. In some cases, data consumers (including the public) are the intended users – that is, the repository designs specific data products or visualizations (e.g., Socioeconomic Data and Applications Center, [\(CIESIN 2024\)](#)). In other cases, a significant part of the community may be data depositors, e.g., researchers individually archiving ad hoc data (e.g., Environmental Data Initiative ([EDI, n.d.](#)), Biological & Chemical Oceanography Data Management Office ([BCO-DMO, n.d.](#))). Many repositories do both (e.g., National Snow and Ice Data Center ([NSIDC 2023; EDI, n.d.](#))).

Repositories are rarely involved in the entire data lifecycle (Ball 2012), typically playing a role during the archival phases (Figure 1, lower). The repository typically assists the scientist in **describing** the data with metadata, becoming responsible for longer term integrity and **preservation**. The repository makes the data available for **discovery** and **integration** into other research efforts, acting as a mediator between data providers (researchers and/or agencies) and all types of reuse. While not optimal, researchers typically submit data to repositories when the project is nearing completion and activities during the planning and collection stages (Figure 1, top right) are often overseen by universities and research funding agencies. In cases where repositories are not involved from project conception, repositories have no control over what happens during data collection; but they can record information about that process during ingestion and can make that information available when data are reused.

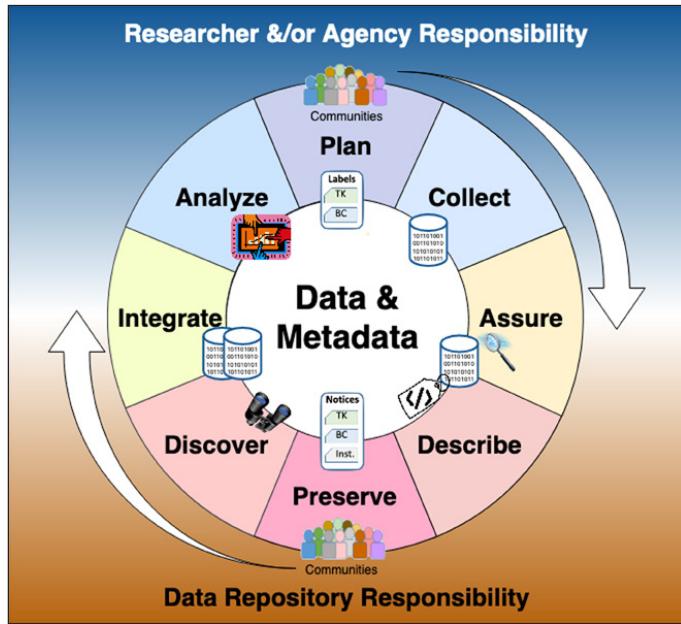


Figure 1 Data lifecycle. Data repositories usually are funded to focus on the stages in the bottom half (orange), in which data are preserved, and not on planning, collection, or analysis stages (blue), which are dominated by researchers and/or agencies. Ideally Indigenous communities would be involved from the very beginning of the planning stage of the data lifecycle. For historical data they may only become involved during data preservation. If a Label exists, it persists throughout all stages of the lifecycle of the data and metadata. See section “Local Contexts” for definitions of Labels and Notices.

CARE PRINCIPLES

The CARE Principles for Indigenous Data Governance state that the application of Indigenous data should result in tangible benefits for Indigenous communities (GIDA 2023a). These benefits can be derived through an in-depth understanding of the community goals and by centering the values-based relationships as defined by community-specific culture and knowledge systems. This push towards inclusively developed outcomes strives to uphold Indigenous communities’ governance, sovereignty, and self-determination. Thus, implementation of the CARE Principles should be required to ensure that the use of data aligns with Indigenous Peoples’ rights (Hudson et al. 2023; Davis 2016).

The four CARE Principles each have three subprinciples (Figure 2). The Collective benefit (‘C’) triad states that data ecosystems must function in ways that enable Indigenous Peoples to derive benefit from the data. Authority to control (‘A’) details how Indigenous Peoples must have the authority and ability to control how aspects of data are represented and identified, e.g., geographical indicators, Indigenous knowledge, or material about Indigenous lands, territories and resources. Responsibility (‘R’) underscores that when working with Indigenous data, there is a responsibility to share how the data are being used and how that supports Indigenous Peoples’ self-determination and collective benefit. Finally, Ethics (‘E’) places the rights and wellbeing of Indigenous Peoples as the focus of concern throughout the data lifecycle and in all data systems.

LOCAL CONTEXTS

Local Contexts is a ‘global initiative that supports Indigenous communities with tools that can reassert cultural authority in heritage collections and data’ (Local Contexts 2024). It introduces the concepts of Metadata Labels and Notices. Labels are applied by Indigenous communities

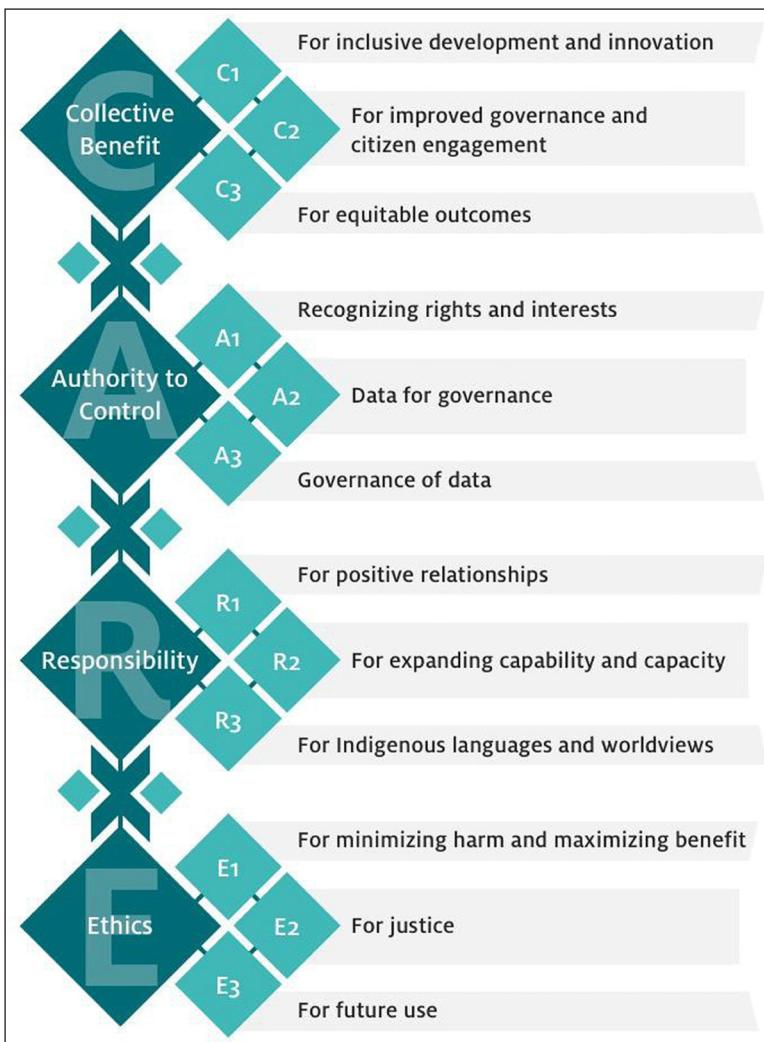


Figure 2 The CARE Principles for Indigenous Data Governance. Reprinted with permission (Carroll et al. 2020).

to develop and assert their conditions for using and sharing knowledge and data. Labels imply that a community has created a rule, which is conveyed publicly through metadata (e.g., data or image). Notices are applied by institutions internally or directly on data, metadata, etc. to recognize rights, and are often used in cases where details are yet to be worked out with the appropriate nations or communities. Traditional Knowledge (TK) Labels (or Notices) cover cultural material that has community-specific conditions for access and use. Biocultural (BC) Labels and Notices cover genetic resources and biological and genomic data. Both Labels and Notices are examples of the broader concept of badges, which in this context involves imagery associated with metadata (Anderson et al. 2013; *Foundation of Digital Badges and Micro-Credentials*, n.d.). For example, data from the recent International Polar Year utilize Polar Information Commons (PIC) badges to 1) indicate the norms and expected behavior of both data users and providers; 2) make the rights of users and providers explicit and 3) to make the data as close as legally possible to open domain (Carlson 2012; Kanao, Okada, and Kadokura 2014).

METHODS

Earth Science Information Partners (ESIP) is a community of data professionals from more than 120 organizations including several US agencies. Their vision is 'a world where data-driven solutions are a reality for all by making Earth Science data actionable by all who need them anytime, anywhere' (Shingledecker et al. 2024). ESIP collaboration areas provide an ideal venue for cross-domain work on common data challenges, and it is in this context that discussion arose about the challenges repositories face when acquiring, curating, and disseminating data and information related to Indigenous Peoples, communities, and their lands.

The Collaboratory for Indigenous Data Governance (hereafter, called the 'Collaboratory') develops research, policy, and practice innovations for Indigenous data sovereignty ('Collaboratory for Indigenous Data Governance' n.d.). They support movements to repair data

collection systems that do not reflect the principles of free, prior, and informed consent, hinder Indigenous access to data, or lack the provenance, permissions, and ethical norms defined by Indigenous Peoples.

All work was conducted during open, online meetings, hosted by either ESIP or the Collaboratory, with an agile, collaborative, and participatory approach in the co-design of knowledge (Noorbergen et al. 2021; Webb et al. 2018). Each meeting was composed of 4–15 people, and overall, approximately 40 from ESIP and 17 from the Collaboratory participated in discussions over 3 years. Indigenous members of this group came from the USA, Canada, Aotearoa (New Zealand), Spain, and Australia; while the ESIP community is based in North America, with some representation from Europe, New Zealand, and Australia. Interim work was presented at larger meetings of ESIP and Collaboratory partners.

Discussions were phased. Phase I began by discussing the need for general guidelines for repositories, especially new ones, on the capabilities needed for FAIR (Wilkinson et al. 2016), TRUST (Lin et al. 2020) and CARE (Carroll et al. 2020) compliance. Lists of repository activities were garnered from sources such as the CoreTrustSeal requirements (CoreTrustSeal Standards and Certification Board 2022), the FAIR maturity model (Bahim et al. 2020) and from ESIP repositories and metadata aggregators such as DataOne (DataONE n.d.) and the Long Term Ecological Research Network (LTER) (LTER, n.d.). After compiling a list of several hundred activities and beginning to map them to FAIR, TRUST, and CARE, the group chose to restrict further discussion only to those activities supportive of the CARE Principles since similar work was occurring in other venues for FAIR and TRUST and repository responsibilities for the CARE Principles was deemed to be a high priority.

Phase II focused on improving the understanding of the CARE Principles by ESIP members, using ESIP meetings with invited Collaboratory members (10–14 hours over 18 months). Collaboratory members provided detailed descriptions of each principle, including background and the ramifications of current practices. As repository managers, ESIP members often attempted to put this new knowledge into the context of a dataset or repository operations that might apply. The result was extensive notes containing a collection of data features or repository activities that were generally a subset of the activities compiled in Phase I. These notes were revised iteratively during subsequent meetings to provide clarification and answer outstanding questions. Also noted were areas of the research cycle in which repositories are generally not involved, such as research planning and review.

Phase II preliminary output was a summary of 30–40 applicable repository activities culled from the discussion notes. The activities were designed to be highly specific, based on experiences with FAIR Principles (Jones, et al 2019; Bahim et al. 2020), because overall, FAIR mappings seem to be most straightforward when features were defined narrowly (e.g., a particular type or use of identifiers) to avoid differing interpretations (Bahim et al. 2020). Some of the activities were recognized as discrete tasks that belonged to a specific area of repositories' operations, such as communication, data curation or technical implementation, which were later revised iteratively. Complex activities were broken down into components. For example, a suggested activity 'attach TK badges to data' was decomposed into specific tasks: communication (to determine which tags belong on which data), metadata needs (e.g., determining if and how current schemas hold notices), and technical implementation (e.g., web display) – essentially the same process as would be used in basic project planning, where tasks are specified for personnel in their individual roles. In practice, these granular activities would eventually be linked together as a project emerged.

The goal of Phase III was the alignment of activities with CARE Principles. ESIP members provided explication of the activities to the Collaboratory. Work was carried out both offline and at online meetings hosted by the Collaboratory with ESIP members invited (approx 10–12 hours over several months). The process allowed for iterative refinement of the activities based on comments and questions. Essentially, this phase reversed the format of Phase II. We should note that this was not planned; however it was clear that we needed considerable Collaboratory input, and so the Collaboratory volunteered a standing meeting, where members' availability could be guaranteed. Using regular, open meetings of both ESIP and the Collaboratory meant that many people could be involved in the discussions who might not otherwise be closely connected to this paper or the outputs. This had the effect of broadening the conversation and recording many viewpoints.

The association of activities to the CARE Principles was led by members of the Collaboratory, although we desired consensus among all participants. In addition to regular online meetings, we recorded opinions offline within a spreadsheet, and revisited decisions at later meetings. This ensured that all members had the opportunity to comment on the final agreed associations. To determine associations, we considered the implications of each activity, mainly who and how an Indigenous community member or their interests would be affected. We did not constrain the number of associations that could be placed on an activity, because precise associations often depended on the type of data to which the activity could apply (see discussion). When an association between an activity and CARE was disputed, the collective opinion of the Collaboratory prevailed. We used a spreadsheet format to collect associations, which was designed to be easily exported as text data. Only final associations are available in the dataset, and comments and meeting notes are available to the collaborators.

RESULTS

Repository activities were organized into four major categories, which represent areas of repository operations and the skill sets of repository personnel (recognizing that skill sets vary widely). Internal communication within the repository bridges the gaps between categories and is not covered here.

- For a repository, *situational awareness* is about understanding the repository environment – including its technical tools, scientific domain, user communities and data holdings, and how these are affected over time or circumstances. On a day-to-day basis, *situational awareness* is the responsibility of data specialists or curators working directly with users. Repository leadership will become involved when significant resources are impacted or for high-level knowledge of communities.
- Outreach personnel are responsible not just for two-way communication with the repository's communities, but also ensure that repository practices and protocols are transparent, justifiable and communicated properly.
- Operational Repository Protocols are generally guided by repository leadership and their policies, with day-to-day procedures typically developed by the staff doing the work.
- Technology is maintained by infrastructure developers, often software engineers who may not have training in a scientific domain.

We identified 47 specific repository activities and organized them into these four categories with no overlap. This list is not meant to cover all possible repository activities, only those identified during Phase II of the project.

The detailed associations between repository activities and individual CARE Principles is shown in [Figures 3–6](#), and summarized in [Table 1](#); raw data are available in tabular form as a dataset ([O'Brien et al. 2024](#)). The entire set of activities and their relationships to the CARE Principles are also available in Zenodo in two forms: a tabloid-sized image (i.e., [Figures 3–6](#), combined) and a printable list ([Duerr et al. 2024](#)).

	SITUATIONAL AWARENESS (11 ACTIVITIES)	OUTREACH (11 ACTIVITIES)	REPOSITORY PROTOCOLS (10 ACTIVITIES)	TECHNOLOGY (16 ACTIVITIES)	TOTAL
C	5	3	4	6	18
A	8	10	6	17	41
R	8	6	6	5	25
E	8	2	11	17	38
	29	21	27	45	

Table 1 Summary of number of associations asserted between the four categories of repository activities and CARE Principles.

The most prominent features are the total number of connections between CARE Principles and repository activities (122), and that most activities are associated with more than one CARE Principle. The most associations between activities and CARE Principles are to Authority to Control (A), followed closely by Ethics (E). Of the four activity categories, the most associations

(and the most activities) were made to activities in the *Technology* category. The sub-principle with the most associations (23) was A1 'Recognizing rights and interests'. Two Ethics sub-principles each had 17 associations (E1, 'For minimizing harm and maximizing benefit' and E3, 'For future use'). There were fewer associations with principles concerned with local (e.g. tribal) governance (A2, C2), capacity building (R2), and ameliorating past injustice (C3, E2), although all these principles had associations to some (five–eight) activities.

RECOMMENDED REPOSITORY ACTIVITIES

The 47 individual activities in each category are listed below, each with a brief description and its relationship to specific CARE Principles and sub-principles indicated by initial and sub-principle number ([Carroll et al. 2020](#)). During discussions, we noted that in some cases, the association between an activity and the CARE Principles would depend on the type of data being managed or other factors related to the activity. For example, the activity 'include funds for specific targeted tools' (R10) would be associated with A3 if the intent for building the tool was to facilitate Indigenous data governance (e.g., a tool to allow Indigenous control over data access), and R2 if the tools were to build capacity in Indigenous communities (e.g., tools for using or displaying the data). In general, we did not deeply consider these alternate factors and tried to associate activities to the CARE Principles that would apply no matter what additional factors were present.

Activities were separated into categories for the purposes of this paper, but it should be clear that an activity in one category will often trigger activities in others, forming a chained response. [Duerr \(2024\)](#) provides a tabular example of chained activities that might be triggered after a repository determines that it holds Indigenous data. The examples in that dataset were extracted from the activities described below.

Situational awareness

On a day-to-day basis and at a fine granularity, *situational awareness* is generally covered by data specialists or curators working directly with users. Such staff typically have a background in a scientific field or area of operations, which attunes them to the needs and priorities of the users with whom they work. At a broader level of granularity, strategic level awareness (i.e. which policy and regulatory frameworks the repository should align to) is generally the responsibility of repository leadership, especially whenever significant resources are impacted or user communities change. Therefore, *situational awareness* activities can cover both strategic activities related to the repository's scope and mission and the day-to-day, data-centric tasks. See [Figure 3](#) for a pictorial mapping between *situational awareness* activities and the CARE Principles.

- **S1 Ongoing engagement with Indigenous communities (C2, A1, R1, R2)**

If a repository becomes aware that they hold or are likely to hold Indigenous data in the future, it is incumbent upon them to engage with the Indigenous Peoples and communities involved (**C2**) in order to develop good working relationships (**R1**), understand the community's rights and needs (**A1**), and carry out more specific activities related to those rights and needs (**R2**).

What this means in practice depends on the repository and the types of data that it holds. For example, a repository co-located with and/or managing data from a narrowly defined location should be able to identify and directly engage with its relevant communities, especially if that community has not already been overwhelmed by requests. An approach to engagement will be very different for a repository that holds a wide variety of global data sets where individual, direct interaction with the roughly 5,000 Indigenous Peoples in the world is simply not feasible (see Discussion).

- **S2 Understand Indigenous legal rights (C3, A1, E2)**

Indigenous Peoples have various forms of underlying customary and legal rights (**A1**). In the US this often comes in the form of treaties or other legal documentation as well as Tribal, Federal or state laws, with rights yet to be recognized articulated in UNDRIP ([United Nations, n.d.-b](#)). It is incumbent on repositories holding Indigenous data to know what these rights are and to make good faith efforts to follow them. This is key to not just recognizing the rights and interests of Indigenous Peoples for new data; but is also the key to identifying and addressing historic injustices (from previously collected data) (**E2**) and for promoting equitable outcomes of data use (**C3**). If a repository is uncertain that a holding complies

with a treaty, law, or rights framework that governed the entities it is or was about, a due-diligence process to clarify the issue should be launched. In the meantime, and depending on the nature of the legal rights in question, distribution of the holding may need to be suspended, its metadata updated to indicate the issue, and careful deliberations on whether it should be deleted or transferred to a different repository initiated.

- **S3 Understand consequences of publishing Indigenous data (E1, E2, E3)**

This is encompassed by the ethics of data publication (**E1, E2, E3**). The fundamental purpose of data repositories is to curate data and information for present and future generations, and to make that data available to that repository's designated communities. However, how and how openly a repository shares data can negatively affect culturally important sites, flora, fauna, as well as people and communities. Many Earth science repositories do not undertake ethical analysis of their operations, creating and publishing a systematic account of the benefits and harms their operations may precipitate. Without this, the reasons for a decision to share or not to share data are often obscure. This activity directs repositories to perform and openly publish an ethical analysis of their publication process, proactively assessing how publishing data may benefit or harm Indigenous communities, including considerations on preventing injustice. This process should include appropriate representation from the Indigenous communities potentially affected, and provide explanation when this is not feasible or possible.

- **S4 Identify types of data of interest to communities (C3, A3, R3)**

As noted in the Introduction, 'Indigenous data' is any data which relates to or impacts Indigenous individuals or communities. Like other types of data categorized in such a manner (e.g. medical data, biological data), the thresholds at which such data requires CARE-aligned contextualization, permissions, oversight, dedicated access conditions (see **S7**), and other measures are hard to identify without guidance from those affected. Further, Indigenous Peoples' data priorities may change over time as their sociopolitical, economic, and physical environment changes (**C3, R3**). To improve identification of indigenous data and application of appropriate data governance (**A3**), repository staff should work collaboratively with communities to identify these data priorities and their changes over time.

- **S5 Understand and respect the changing needs of communities (R1)**

Indigenous Peoples' and community needs for any given dataset might change over time. It cannot be understated that those needs should be understood and respected at the time of deposition, but since change is inevitable, the rights and interest of the communities should remain a focus into the future. Part of this change will be establishing and adapting engagement mechanisms (see S1, above) with Indigenous Peoples (**R1**) in sustainable ways.

- **S6 Be aware of changing roles/relationships over time (R1)**

Repositories need to be prepared for continuously changing roles. Indigenous Peoples' needs and priorities will shift over time as the global context changes, people age, or new technology becomes available. Recognizing and adapting to change is important to maintaining positive relationships with Indigenous communities (**R1**). For example, if an agreement with a community currently specifies that a particular person is responsible for approving data access requests, it needs to be understood that who that person is will change over the years and that even the approval mechanism itself will change over time. Thus, repositories should ensure that a periodic review of roles and relationships is performed, to ensure clarity and reduce misunderstandings.

- **S7 Determine if data access restrictions are necessary (C3, A3, R3, E1)**

For each dataset held or expected, the repository needs to leverage the previous *situational awareness* approaches (S1–6) to determine whether any part of the data needs to be restricted access or have usage restrictions communicated in its metadata (**A3, E1**). Indigenous community protocols may require data to be restricted to members of a particular sex, family, or during certain times of year (**R3**). Further, a community may wish to specify who is responsible for approving data access requests. Access restrictions support principle **C3**, and promote Indigenous worldviews and ethical frameworks.

- **S8 Determine if data obfuscation is necessary (A3, R1, R3, E1)**

Depending on the data content there may be a legal or community preference for all or certain parts of a data set to be obfuscated (**A3, R1**), such as to change the accuracy of the location of sacred sites or the ages of people so that the data are useful to end

users, while any sensitive information is hidden (**R3, E1**). This is similar to what is often recommended in biodiversity studies, when handling endangered species occurrences ([Chapman 2020](#)) (and coincidentally, data that is also likely to be of interest to Indigenous Peoples). Practices for identifying these will be similar to determining if restrictions are necessary.

- **S9 Recognize when collection agreements were needed for data (C3, A1, A3)**

Data collection agreements are generally made between researchers and (appropriately empowered representatives of) Indigenous communities, and are mechanisms for operationalizing rights and interests (**A1**). These agreements typically stipulate that community-sensitive collection and stewardship protocols are followed (**A3**), and are intended to ensure more equitable outcomes when the data are used (**C3**). Requirements for data collection can be highly context dependent and may be embedded within a more encompassing research plan or proposal. Collection agreements may have legal force from the local to global levels, or may simply be promulgated by individual communities as expected researcher practices or negotiated research relationships ([ITK and NRI 2007](#)).

For data without agreements that are already in the repository (or that pre-date agreements), the question becomes what actions should be taken to clarify the status. Solutions could range from working with the appropriate Indigenous community or communities to create retroactive agreements, infrastructure to support proper re-curation of the data, to returning the data to the Indigenous community (see discussion).

For new data, a repository should ensure that a depositing researcher has done their due diligence and that the data are accompanied by any needed documentation. If this work has not been carried out, data may be rejected, or the repository may have a mechanism to work with the affected community to remedy the situation.

- **S10 Ensure design of data products, metadata and tools are appropriate for communities (R3, C2, A2)**

Our definition of Indigenous data includes ‘data that relates to Indigenous Peoples’, which can be very broad. For some data, issues are not about Indigenous control, but instead are about ensuring access to the data (e.g., weather data) in ways that are understandable and usable (e.g., language, accessible formats, proper metadata) ([Taitingfong et al. 2023](#)) (**R3**), and that can improve tribal governance frameworks and engage their community (**C2, A2**). Relevant communities should be engaged and working with the repository to define the products they need rather than the repository making assumptions.

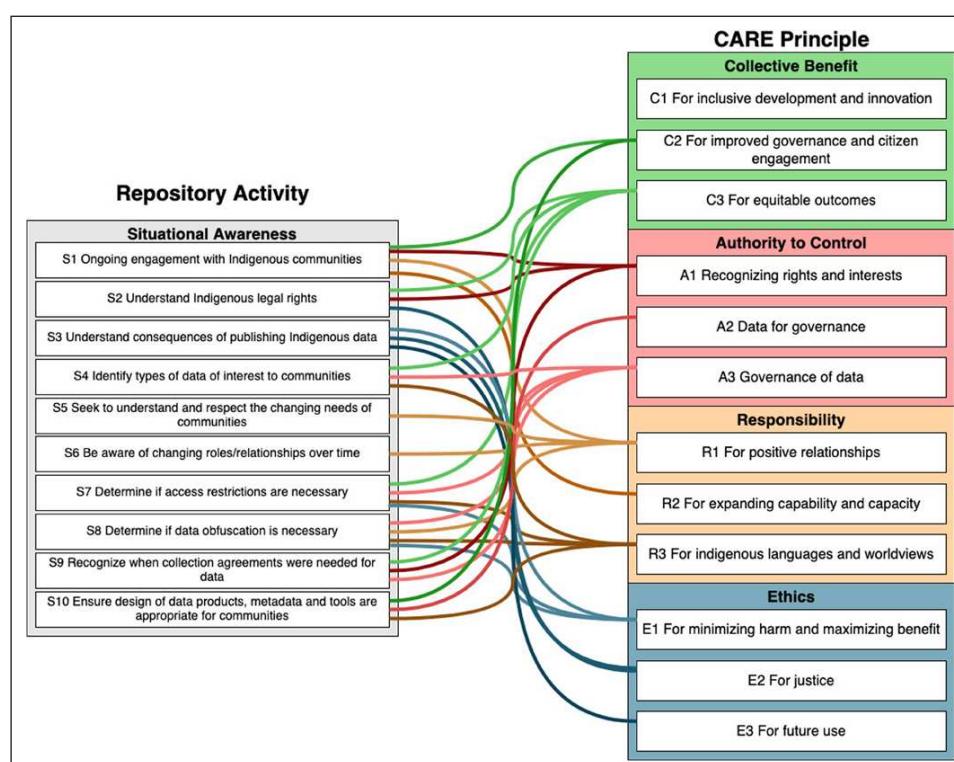


Figure 3 Associations between activities classified as *situational awareness* and CARE Principles.

Outreach activities entail not just direct communication with the repository's communities, but also in ensuring that repository practices and protocols are transparent, justifiable, and communicated through repository systems. Despite its importance, this latter function is often secondary to promotional outreach. Some repositories have a designated officer or user services group who handle user queries, contact lists, web page construction, seminar schedules, etc; others approach communication and outreach ad hoc. Outreach activities are generally concerned with a) communicating with repository users, b) engaging new communities or sub-groups. See [Figure 4](#) for a pictorial mapping between outreach activities and the CARE Principles.

- **O1 Facilitate relationships between data provider and user communities (R1)**

Most repositories are intermediaries working with potential data providers to ingest, preserve, and to make that data available for end-users ([Figure 1](#)). They are uniquely positioned to facilitate relationships (**R1**), to ensure Indigenous rights and perspectives are taken into account when appropriate. This can only occur if the repository is involved with projects at early stages (e.g., a research proposal) rather than as a last step (e.g., publishing research).

- **O2 Include Indigenous representatives on repository advisory board, if relevant (A1, A2, E2)**

Repositories holding Indigenous data should include appropriate Indigenous representatives on their external advisory boards to ensure awareness of Indigenous rights, interests and needs (**A1, A2**), while addressing any past or potential ethical imbalances (**E2**) and ensuring appropriate engagement mechanisms. Implementation will vary widely depending on the size and scale of the repository and its data holdings. For example, the current advisory board of Exchange for Observations and Local Knowledge of the Arctic (ELOKA) program at NSIDC consists entirely of Indigenous representatives because ELOKA works solely with Indigenous communities on the 'collection, preservation, exchange, and use of local observations and Indigenous Knowledge of the Arctic' ([ELOKA 2023](#), emphasis added). The advisory board for a repository containing more global datasets supporting diverse, multidisciplinary user communities will not be able to have representation from all communities its holdings may impact or interest while still being maintained at an effective size, and would benefit from referring to external bodies capable of managing focus groups, surveys, etc. (see Discussion).

- **O3 Anticipate working with more communities (C2, R1)**

Repositories should be prepared to engage with communities that were not in the original scope of the activity. Preparing a repository to be able to work with new communities (**C2**), for example those with different language needs, takes effort and may involve adding staff with the needed skills. In some cases, simply accepting a data deposit may require adding new designated communities, including Indigenous communities ([Baker Duerr and Parsons 2015](#)). Repositories should identify and reach out to potentially interested parties (**R1**).

- **O4 Ensure repository practices are publicized and transparent (A1)**

The detailed day-to-day practices of the repository should be public and clearly stated (e.g., data deposit, data management, etc.), except where this might compromise the security of the data. This can help demonstrate recognition of Indigenous rights and interests (**A1**) and is essential to establishing trusted relationships.

- **O5 Publicize your governance protocols (A1)**

Repository governance protocols and policies need to be written in forms that are understandable by your designated communities. For example, this might include translation into preferred languages. These should be made publicly available through all existing communications channels. The purpose is to facilitate Indigenous communities' understanding of your repository's capabilities and processes to inform their decisions about deposit and access to data you hold, e.g., a decision about what Indigenous data would be appropriate for your repository to manage and further demonstrates a respect for their rights and interests (**A1**).

- **O6 Identify your repository as a holder of Indigenous data (A1, A2, A3)**

Transparent and clear communications about any Indigenous data holdings is key to enabling Indigenous data access and data governance (**A1, A2, A3**). Websites, brochures,

videos and other material should indicate whether data restrictions, badging, or other mechanisms are being used. For some, e.g., global datasets, this will be complex (see discussion).

- **O7 Liaise between communities and vocabulary maintainers (A1, R3)**

Repositories are rarely directly involved in the development of all of the vocabularies (thesauri, ontologies, etc.) associated with the data they contain, and such connections are even more tenuous for researchers or users, such as those from Indigenous communities (**A1**). However, many vocabulary maintenance organizations have mechanisms for public input (e.g., issue trackers, public review periods). Participate in these processes to ensure that vocabularies include the terms, definitions and languages (**R3**) needed.

- **O8 Share materials with the Indigenous workforce (C1, A2, R2)**

The Indigenous materials curated by the repository (e.g., digital data, samples) need to be available to the Indigenous Peoples to which they relate. These products are essential to facilitating capacity building and innovation within the community (**R2**), enabling self-determination and inclusive development (**C1**) and to promote local governance (**A2**). Outreach activities should therefore co-develop effective methods to share materials with interested Indigenous communities.

- **O9 Clarify types of uses with communities (A3, E3)**

Another topic for discussion with Indigenous Peoples is the community's preferences for possible data restrictions and acceptable uses, to ensure Indigenous control where appropriate (**A3**). This information should be included in the metadata associated with that data (e.g., information about associated Traditional Knowledge Labels ([Local Contexts, n.d.](#))) and may result in data obfuscation or embargos (**E3**)).

- **O10 Identify educational opportunities with communities (R2)**

One of the topics that should be discussed when working with Indigenous communities is what types of educational activities would be welcome and useful, to promote Indigenous capabilities and capacity (**R2**). This information can be used in equitable prioritization of any existing educational funding or for pursuing additional funding.

- **O11 Identify types of data that cannot be supported technically (C1, R3)**

Some of the capabilities required to fully support the CARE Principles may not exist within a repository's current infrastructure. Understanding and being open about a repository's capabilities – for example to support data embargos – is necessary to ensure that community expectations about data representation, acquisition and access are met (**C1, R3**).

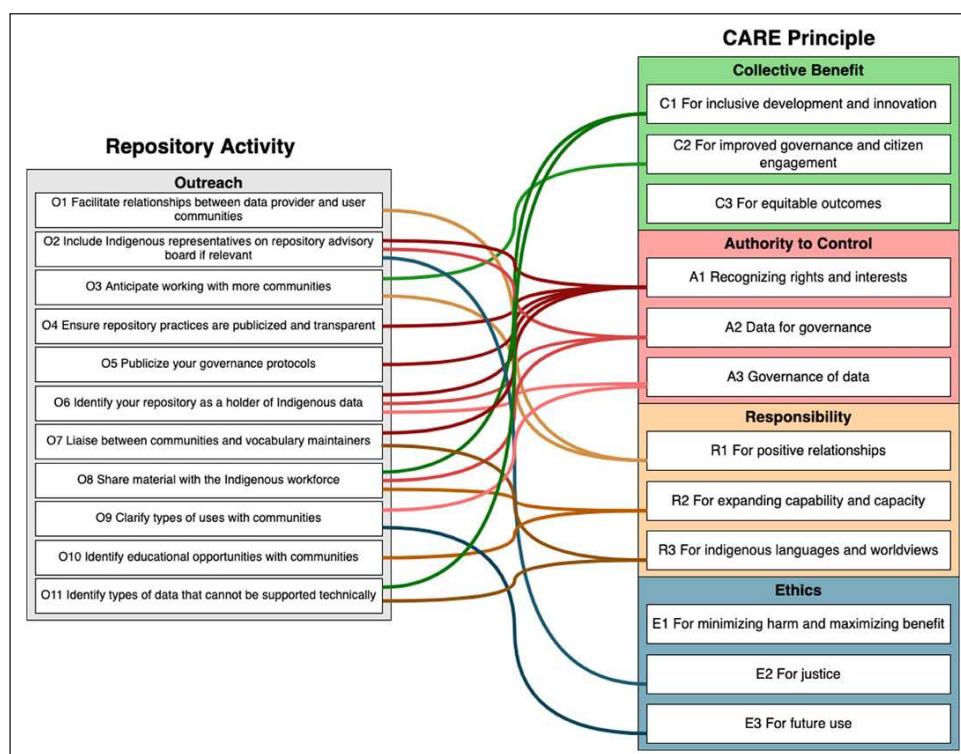


Figure 4 Associations between activities classified as Outreach and CARE Principles.

Operations are generally guided by repository leadership, with day-to-day procedures typically developed by the staff doing the work (e.g. data curators). Implementing CARE will impact both high level policies (e.g., choosing holdings) and routine operational procedures (e.g., user contact). An effective repository develops their protocols – including both policies and procedures – with key user groups in mind, which means taking into account Indigenous community knowledge and concerns. Recognizing Indigenous rights and interests and incorporating their input promotes improved engagement and decision-making, resulting in stronger relationships, increased trust, and reduced harm. By placing Indigenous community members in a position to shape institutional policies and workflows, a repository can promote improved Indigenous governance, ethics, and equity in their collections. See [Figure 5](#) for a pictorial mapping between repository protocol activities and the CARE Principles.

- **R1 Develop data policies and acquisition/curation procedures (A1, R2, E1)**

A repository's data acquisition policies advertise the types of data within the repository's scope and also types that cannot be handled due to technical, legal, or other limitations. These policies should explicitly describe their effect on Indigenous Peoples and data. Acquisition procedures will benefit from requiring steps to determine 1) if the data proposed for inclusion is Indigenous data (**A1**), 2) what Indigenous communities should have been involved in its acquisition, 3) which (if any) agreements apply, 4) potential future use (**R2**), and 5) what actions the repository will take if the appropriate agreements were not in place at the time of data acquisition (**E1**). Further actions might include working with the Indigenous community on an agreement, or rejecting the data deposit. Data acquisition procedures should also include an investigation of the content of the data to determine potential restrictions or metadata needs (see the *situational awareness* section).

- **R2 Ensure data management policies are justifiable (E2, E3)**

Well-constructed policies should have clear justification and supply context for decisions. Policies should illustrate a repository's awareness of legal and ethical issues, and are evidence of processes that positively impact the creation, curation and use of digital objects ([CoreTrustSeal Standards and Certification Board 2022](#)). External guidance from Indigenous communities is particularly important to ensure adherence to ethical principles and regulations (**E2, E3**). Justification of standing policies becomes especially important when conflicts arise or when the funding is inadequate to cover a request, such as a technical improvement.

- **R3 Develop procedures to work with depositors to address necessary restrictions (A1, E1)**

When acquiring Indigenous data where restrictions are necessary, the repository should work with the depositor to address these (**A1, E3**). Activities might include 1) obfuscating part or all of the data to protect restricted material; 2) segregating the data into restricted and unrestricted portions with separate access processes for each and/or 3) removing sensitive data from the deposit. Ideally the depositor will have structured the data accordingly.

- **R4 Develop procedures to request required documents during data submission (A1, E1)**

The repository must be aware of acquisition or other agreements that apply to data being deposited, to ensure that Indigenous rights were recognized (**A1**). Without ethical procedures, reuse may be harmful (at worst) and even at best, is not likely to maximize benefit to the Indigenous community (**E1**). This is, of course, contingent on the repository being aware that such agreements were needed. As this is often not the case, a passive mechanism requesting submitters to verify that all such agreements are in place (expressly identifying those pertaining to Indigenous Peoples) could be deployed.

- **R5 Develop procedures for determining if needed technologies are within technical plans (E1, E3)**

Every repository has a vision of what can be accomplished technically given their current budget. In implementing CARE, it is likely that data restrictions or special curation activities will be required. These might include additional metadata, application of badges such as TK Labels, which are identifiers for cultural material that has community-specific

conditions regarding access and use (Local Contexts, n.d.) (E1). Procedures should be in place to determine if this can be supported by current technical plans. If not, repositories should have a mechanism for future planning (E3). If procedures are transparent (see the Outreach section), both depositors and relevant communities will have the knowledge to make informed decisions.

- R6 Develop processes to identify communities for engagement and collaboration (C1, E1, E2)**

Repositories should develop procedures for determining communities likely to have an interest in the types of data they hold. Mechanisms could include assessing the geographic areas of data deposited, using contextual information in the data to identify associated Indigenous Peoples and communities, and proactively taking into account Indigenous communities in repository future plans (E1, E2). Additionally, many repositories help researchers develop data management plans for proposals, and this engagement may identify potential future communities for collaboration prior to data acquisition or even funding (C1).

- R7 Develop community contact protocols and follow up processes (C1, A1, R1, E3)**

A repository's protocol for an initial contact with an Indigenous community should begin with using existing resources (e.g., websites, books, etc.) to learn about that community and how it operates. For example, many Indigenous Peoples have dealt with researcher and other requests for years and have well established contact protocols (ITK and NRI 2007). Following such an established protocol is respectful when attempting to enter into a new relationship with a community (C1). In the absence of such protocols, repositories or better yet consortia of repositories should reach out to identified community groups or members with a brief description of the Indigenous data of interest. The goal being to 1) assess whether there is community interest in collaboration and if so, 2) to identify the proper communications paths (A1, R1). Such contact protocols can be used both for when you determine that you already hold Indigenous data and for accepting new data deposits (E3). Once contact is established, procedures for when and how to follow up with the community should be negotiated (C1, A1, R1).

- R8 Develop continuity plans (A2, R1, E3)**

It is a repository's responsibility to ensure the longevity of the data in their care (E3), no matter the circumstances, up to and including dissolution of the repository (CoreTrustSeal Standards and Certification Board 2022). Beyond simple backup and recovery strategies, a repository's plans for handling Indigenous data in its care should include consultation with the affected communities (A2, R1). In the most complex cases this may result in transfer of data to another repository or repatriation of the data back to the community.

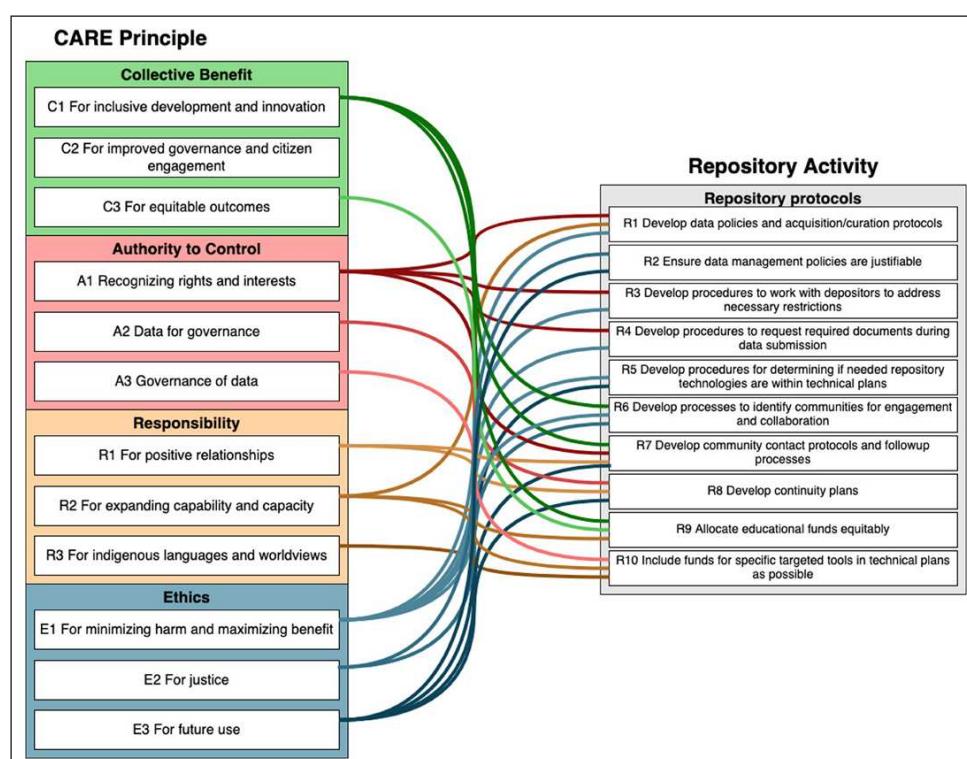


Figure 5 Mappings between activities classified under Repository Protocols and CARE Principles.

- **R9 Allocate educational funds equitably (C1, C3, R2)**

Any general educational activity funding should be spent equitably (**C1, C3**), ideally based on the educational activities identified through communication (**R2**) (above). It should be noted that in many cases, funding for educational programs is not in a repository's base funding and must be pursued separately.

- **R10 Include funds for specific targeted tools in technical plans as possible (A3, R2, R3)**

If needs for specific targeted tools were identified during data ingest consultations with communities and funding for development is currently available, equitably include funding of those tools in your technical plans (**A3, R2, R3**). If funding is not available, consider applying for supplementary funding for their development.

Technology Development and Maintenance

Technical infrastructure is maintained by developers, who are usually software engineers or personnel with expertise that might be limited to fields like web or database development, computer science, or network hardware. Generally, activities here are concerned with the creation and maintenance of the underlying systems required for almost all repository operations. User-facing components are often software, so technical features are how most people interact with the repository (e.g., through its website), and how policies are implemented (e.g., a decision to require logins). Here, our goal is to guide such implementations to enhance alignment with the CARE Principles (Figure 6).

- **T1 Enable granular embargoes (A1, A3)**

Embargoes restrict access to entire datasets or their parts, and may include a time limit. This could include requiring community approval before providing access via a restricted account (see for example the Gwich'in Place Names Atlas ([Gilbert et al. 2017](#)) where a request placed to NSIDC's user services is forwarded to the community for approval). Alternatively, secure systems such as those used by social science archives like the Inter-university Consortium for Political and Social Research (ICPSR) for highly sensitive data could be established ([ICPSR 2024](#)). Specifics will depend on the context and content of the data, and needs of the community (**A1, A3**), but implementation of secure and user authentication and management systems and encrypted and protected storage are advised.

- **T2 Support tracking data provenance (A2, E2, E3)**

Data provenance can be viewed in multiple ways, e.g., to indicate a source (where does the data come from), reuse or derivations (where has it gone), and changes (what has happened to it along the way) (**A2**). Tracking reuse can enable concerned communities to verify that agreements are being met and ensure that the data's value is accruing to those communities' wishes (**E2, E3**). The Indigenous Data Working Group (IDWG), through the P2890 project, is drafting an IEEE Recommended Practice for the Provenance of Indigenous Peoples' Data ([GIDA 2023b](#)) which is expected to be released in 2025, and can further inform repository practices. We thus recommend that repositories implement or enhance their existing provenance tracking systems (e.g. using W3C PROV standards) in preparation for harmonization with recommendations and needs from Indigenous communities. This entails granular information management which is closely tied to repository operations and will require coordination with other upstream or downstream repositories or data aggregators.

- **T3 Link between data, papers and other outcomes (A2)**

Providing links between the data deposited in repositories and their use outside the repositories, such as in derived datasets, papers, patents or tools helps promote self-determination and governance (**A2**). It also helps to verify that agreements are being met and that the value of the data can accrue to the community, not just those that acquired or used the data. While repositories cannot control whether downstream users will actually link back to their holdings, they can implement measures to make such links as easy to implement and technically useful (i.e. actionable) as possible. For example, implementers may ensure all holdings have de-referenceable permanent identifiers (PIPs, URI/IRIs) in URL form and guidance (embedded in metadata) on how to cite the resource such that external use is detectable using standard approaches (i.e. ping logging services, web crawling, URL tracking).

- **T4 Create data access control mechanisms (who/when/changing roles) (A1)**

Mechanisms for recording and controlling who can (and did) access data and when it can be accessed help ensure that the rights and interests of the relevant Indigenous Peoples are taken into account before data is released and 2) that agreements and commitments made by those accessing the data are upheld (A1). This includes potentially allowing for external review and approval of access requests and keeping track of who and how this changes over time (see S6, S7). Appropriate technical measures may resemble those described in T1, but require immutable, version-controlled logs of access authorizations and agreements, and possibly distributed ownership (e.g., copies residing with the repository, the access-granting authority, and a neutral, mutually trusted third party). Token-based exchange systems such as Open Authorization (OAuth) may be explored for more granular or dynamic exchanges, if implementation partners have suitable capacity.

- **T5 Implement systems that allow transfer of responsibilities (C2, A1, A2, A3, E1, E3)**

As has been previously noted (see Protocols), it is a repository's responsibility to ensure the longevity of the data in their care. This includes the ability to transfer responsibility for Indigenous data and related tools and technologies (A1, A3) either to the associated community or to a third-party designated or agreed to by that community (E1, E3, A2). Such responsibility transfers can occur for many reasons (e.g., loss of repository funding, changing Indigenous community needs). Such responsibility changes can support improved Indigenous data governance (C2, A3). This transfer – due to the other technical requirements of handling Indigenous data – can be a complex task, and may require too many changes in the receiving system to be viable. Thus, technical modularity, containerization, and the use of other transferable system components (e.g. cloud-based storage and compute modules) should be explored.

- **T6 Package data with associated, required documents (A1, E1, E3)**

Metadata specifications and packaging protocols should support associating contextual information permanently, in order to protect the rights and interests of Indigenous Peoples through time (A1, E1, E3). Documents such as agreements, permits, extended descriptions, and histories should either be contained in a data distribution or de-referencable, permanent identifiers (e.g. DOIs) which resolve to securely archived copies of these documents should be included in the data's metadata. When external document repositories are used, these must be well-vetted to ensure they have mature and reliable archiving and service models.

- **T7 Support communication of rights and restrictions (A1, E1)**

Repository discovery and access systems should provide metadata about associated rights and restrictions even when data are not immediately available. This allows communities to verify that the agreements that have been reached are being met. Per O9 above, the form of this communication (e.g. the properties used in metadata packages) should be validated and agreed upon with the relevant Indigenous communities.

- **T8 Implement badging mechanisms (E1, E3)**

Repository systems and metadata schemas should be able to associate Traditional Knowledge (e.g., [TK Labels – Local Contexts](#) n.d.), BioCultural Labels ([“BC Provenance – Local Contexts,”](#) n.d.) or other forms of badges in order to permanently express the relationship between Indigenous communities and their data. This helps ensure ethical future use that maximizes benefits and minimizes harm (E1, E3).

- **T9 Use a metadata format that handles multiple languages (C1, A1, R3)**

English has become the lingua franca for science. However, in order to support inclusive local development (C1), Indigenous data governance and worldviews (R3) and to make data accessible and understandable to a range of communities (A1), it may be necessary to express metadata, including metadata within the data itself, in multiple languages including Indigenous languages. This may require the use of new character sets and encodings, and thus configuration changes throughout data handling routines.

- **T10 Keep a list of data “actors” and be able to link data to them (A1, E2, E3)**

Repository systems must be able to associate designated Indigenous individuals and organizations with rights and responsibilities that affect the operation of the system (A1), such as the right to approve access requests (E2). Responsible parties may change over time, which should be recorded as part of the provenance of the data (E3). Closely tied to T1, T2, and T4, repositories should maintain an actionable list of actors (and

metadata about them), using reliable and permanent identifiers. Ideally, the actors will maintain their own identities in a mutually-acceptable system with content harvested by the repository.

- **T11 Use community-vetted vocabularies (A1, C1, R3)**

Data that comes from a community should be understandable and usable by that community (Eitzel et al. 2017). Well-maintained vocabularies (here, inclusive of thesauri, ontologies, and other semantic resources) mean that all or part of the data, its documentation and metadata used terms relevant and understandable to Indigenous communities (A1, C1). Expressing terms in multiple languages is an option for sophisticated vocabularies (R3), and definitions and comments to clarify proper interpretation and nuance are essential. For example, the website of the Clyde River Weather network (ITTAQ 2024), originally established as part of a joint project between local hunters and Elders in Nunavut and researchers from the University of Colorado Boulder and Colorado State University under the auspices of ELOKA is available in both English and Inuktitut.

- **T12 Support customized data views (C1, C2, R3)**

Engagement (C1, C2) and decision-making is improved when a community can access data through meaningful views (R3). This may require the development of data derivatives and custom visualizations. If a repository does not have the capacity to develop these themselves, they may need to provide access routes (e.g., secured application programming interfaces (APIs)) to third parties trusted by communities to provide these services (see discussion).

- **T13 Identify metadata fields for recording required documents, proper use and limitations (A1, E1, E3)**

Development of an Indigenous metadata bundle to guide placement of associated material in metadata is currently in progress (Taitingfong et al. 2023). In the interim, repositories and metadata exchange systems should include either links to or verbatim usage restrictions and guidance, licenses, documentation, and related information pertaining to Indigenous communities in the same fields used for their other users (e.g. publishingPrinciples, license, usageInfo in schema.org metadata). Ensuring that this metadata is present cannot prevent abuse; however, when placed in fields that repositories and other systems parse in due diligence processes, it will make information that clarifies the rights and interests of Indigenous communities more discoverable and actionable (A1), and thus increase capabilities to minimize potential harm and maximize benefits (E1, E3) throughout the lifetime of the metadata.

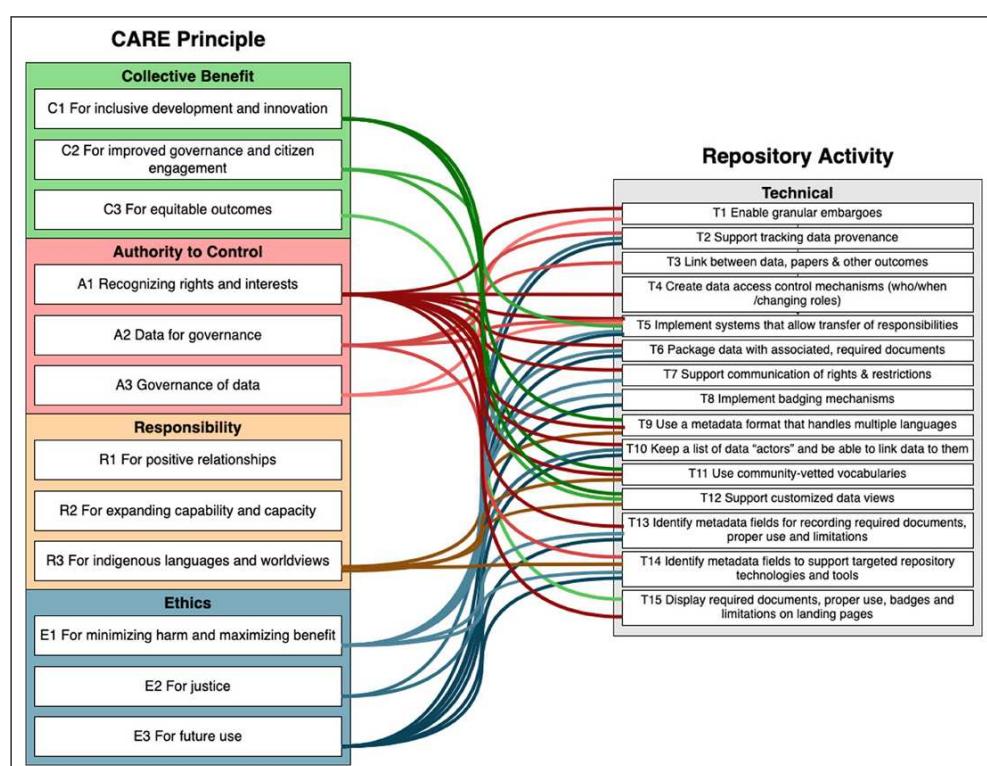


Figure 6 Mappings between activities classified under *Technology* and CARE Principles.

- **T14 Identify metadata fields to support targeted repository technologies and tools**

(A2, R3, E1, E3)

Supporting repository technologies and tools designed specifically to support Indigenous community data needs (**A2, E1**) often requires additional metadata aligned to Indigenous languages, worldviews and values (**R3**). When technologies and tools have been created for a dataset, the repository should link between these and the data to ensure they or their equivalents remain available for the long-term (**E3**). To date, ESIP has not identified any metadata standards which contain clear fields for these types of metadata (see Discussion).

- **T15 Display required documents, badges, proper use and limitations on landing pages**

(A1, C3)

User interface systems for displaying data and data summaries need to be able to effectively display and highlight information about the required data. Indigenous (meta) data is no different: repositories should thus ensure that (meta)data such as TK Labels and badges, proper uses and limitations of Indigenous data are visualized clearly to users during their interaction with repository services. This makes the rights and interests of Indigenous communities visible to the larger community (**A1**) and when combined with appropriate protections (above) helps promote equitable outcomes (**C3**).

DISCUSSION

Because of the high diversity among the size, missions, governance models, staffing and domain specificity of the more than 3000 data repositories in the world (per re3data.org), and the high diversity of Indigenous Peoples, these recommendations are intended only to start and help structure conversations among repositories considering steps toward CARE implementation. Some repositories may already have or be progressing towards a technical basis that can support the implementation of the Technical activities described above. However, many repositories have limited engagement with users aside from technical support and it would be rare for a repository's external advisory board to cover all interested communities.

To align their operations to the CARE Principles, repositories will need to make a number of changes. Some of the changes are internally focused on *Technology* and *Repository Protocols* that make repositories more responsive to Indigenous data. Others will need to occur in the area of communications, in particular, *Outreach* – the engagement of communities, and *situational awareness* – e.g., a comprehensive understanding of all aspects of a data set to ensure the best outcome for an Indigenous community. Naturally, any data application targeted for a particular community should have gathered input from that community during the design phase. When an Indigenous community is the intended beneficiary of such an application, this is especially important to avoid (typically inadvertently) perpetuating colonial legacies. Indeed, given the complex and nuanced nature of Indigenous data – and its relations to traditional knowledge systems and socio-political sensitivities – the consultation with relevant Indigenous communities during implementation of the CARE Principles is essential to fully realize alignment to their intended meaning.

OBSERVATIONS

First, we note that this work was not an empirical survey, but is the result of discussions among approximately 60 people over several years. Consequently, it is important to temper interpretations of our results; as noted above, this work is meant to initiate further conversations rather than provide authoritative guidance.

In the results, we noted nearly all repository activities were related to more than one CARE Principle, which is consistent with other work linking CARE to data stewardship practices ([Belarde-Lewis et al. 2024](#)) and the interdependence of the CARE Principles themselves. The most associations made were to activities in the *Technology* category, which reflects the interdependence of repository activities within the ESIP community and that technical infrastructure supports the other three activity categories. We noted that there were fewer associations with principles concerned with local (e.g. tribal) governance (**A2, C2**), capacity building (R2), and ameliorating past injustice (**C3, E2**), although all these principles had associations to some (five-eight) activities. These trends are logical, since repositories are focused on data preservation, access and support of data's future use ([Lin et al. 2020](#)), and

generally speaking, recognition of rights and allocation of benefits can be communicated through dataset metadata. The CARE Principles that have lower numbers of alignments also might reflect the fact that those needs are likely to have more complex solutions than can be accomplished within a data repository.

Activities that represented multiple steps of a process were broken down (e.g., as described in Methods for 'attach TK Labels to data'). However, in an effort to simplify, we also combined what initially appeared to be some very similar activities. For example, an early iteration contained separate activities pertaining to needed metadata fields (for uses, permits limitations, badges, other documentation). For simplicity, these were combined into one, T13, ('identify metadata fields for...'). However, those needs (permits, badges, etc.) are somewhat disparate, and each is related to a different CARE principle. Therefore, T13 has many associations to CARE, whereas simpler alignments might have been achieved if they remained separated. Likewise, a multistep activity like 'apply TK badges to data' alone would probably be aligned primarily with E3 (ethics of future use), but its components steps are aligned to many principles as they cover communication, intended use, and technology. In a project management context, using the multistep examples of the Duerr (2024) dataset may result in fewer and more streamlined associations.

ACCESS AND CONTROL

The definition of 'Indigenous data' used here is quite broad, and could be argued to apply to any data, anywhere. For example, a concern arose in the ESIP community that, with a definition this broad, even something like the daily weather would count as 'Indigenous'. Certainly these data should be included – especially as climate change causes more extreme weather – making related data available to Indigenous communities in ways that are easily accessible and understandable can be literally a matter of life or death.

The issues of handling (and solutions) will differ for different types of data. For example, for climate data or general biodiversity data, access may be the major concern. Whereas, for data from or explicitly about Indigenous people, the main issue is likely to be one of control. In this discussion, the question of control by Indigenous communities is considered separately from access, since access is a primary function of many repositories, and some issues of control may be beyond the scope of some repositories to resolve (see below).

Access to data in ways that are meaningful and useful is a fundamental requirement for any user. For some, simple permission settings might be enough. However, products tailored to community-specific needs may be important for some Indigenous Peoples, and logically, must be designed with their input (e.g., as described in Christen 2012). As noted above, granular, well-structured, -annotated (meta)data and derived products will support user interface/user experience developers in creating such tailored views. Vis a vis the weather data mentioned above, solutions might include tailoring warnings, forecasts, and other weather products for users with low bandwidth, or to provide products in Indigenous languages.

Sovereignty and control become relevant when data are explicitly about Indigenous Peoples or their territories, whether generated by them or by others (Carroll et al. 2020). In these cases, data exposure or reuse may need to be controlled by the Indigenous community or its designate. Data categories could be developed to help prioritize repository work, ideally in conjunction with the Indigenous communities involved. For example, a high priority might be to identify data collected manually on Indigenous land or about Indigenous Peoples or artifacts, where sovereignty is clear, permits were required, and communication is relatively straightforward (e.g., tribal elders can be identified). The Smithsonian instituted Directive 609 (Smithsonian Institution 2019) to recognize the complexity within their digital assets and metadata, which identifies specific restriction categories for material that might need additional management, with general guidelines for documenting the nature of the restriction in collection systems. The directive also outlines which roles within the institution are responsible for aspects of improved stewardship and potential restrictions.

However, it should be noted that once data is publicly available, knowing how it is being used becomes practically impossible. Two possible remedies for that would be to impose access protocols that include legally binding contractual or usage constraints or to add information

(i.e., metadata) to the data that describe the permissions, protocols and usage constraints that should be used. In either case, the problem becomes one of creating tracking and verification mechanisms as well as determining what could be done if violations are detected. All of this is beyond the scope of the work done here.

Lastly, as mentioned in the recommendations, in some cases returning the data to the Indigenous community may be necessary (e.g., upon loss of repository funding or when data repatriation is needed). It needs to be recognized that undertaking such a transition is likely to be quite complex, involving at the very least the transfer of the data and its metadata to the community, deleting instances of the data from all servers and backups under the control of the repository, replacing the targets of identifiers for the data and its metadata with notices recording this action (unless requested not to), and supporting propagation of these changes to all data and metadata aggregators, such as DataCite ([Vierkant, n.d.](#)), that maintain records about the data.

CALLS FOR OPEN DATA

The CARE Principles are designed to complement FAIR, TRUST and other mainstream data frameworks ([Wilkinson et al. 2016](#)) ([Lin et al. 2020](#)). FAIR is focused mainly on data clarification and reusability plus simplified retrieval, while TRUST stresses repository operational issues like transparency and sustainability – tenets designed to promote global data openness and awareness. CARE extends awareness to promote equitable participation and outcomes from said data access, use, reuse, or attribution, and emphasizes that the level of openness and accessibility be aligned with Indigenous rights. Resolving the tension between protecting Indigenous individual and collective rights and interests while encouraging open data in a global research environment can only be accomplished by strong ties between the repositories entrusted with that data and the Indigenous Peoples and communities that relate to that data, in particular, by increased communication and adoption of Indigenous standards. Yet, as noted below that communication is itself fraught.

A fundamental purpose of data repositories is to curate data and information for future generations and to make that data available to the repository's designated communities now and into the future. With current increasing calls for open data, funders and repositories may be assuming that all data can be '... freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike' ([OKFN 2016](#)). In reality, even the full definition from the Open Knowledge Foundation recognized the need to preserve provenance (version 2.1, *ibid*), and further, that early conversations around data openness and sharing were incomplete ([Avila 2023](#)). In fact, most open data calls discuss the concept of ethically open data, because unrestricted public access or 'open by default' might be incompatible with the need for authority, ownership and control. Data that align with the CARE Principles is ethically open data ([Kukutai 2023](#)).

IMPLEMENTING CARE: REPOSITORIES CANNOT DO IT ALONE

Repositories have an important role to play in implementing the CARE Principles, but they cannot resolve all of the issues or tensions that exist. Most of the issues below will require discussions among not just repositories but governments, international bodies and funding agencies, in collaboration with consortia of Indigenous Peoples who have specific interests in the outcome.

First, are issues of scale. Controlled access may work well for isolated data about particular people or places that are rarely studied, but controlled access on a granular level becomes almost impossible for data collected at large scale, e.g., global data from satellite remote sensing. Worldwide, Indigenous Peoples represent at least 5,000 distinct groups ([United Nations, n.d.-a](#)), thus it is unrealistic for a single repository or even group of repositories to consult and collaborate with all. Similarly, it is unlikely that any Indigenous Peoples would have the resources to consult with all of the many thousand repositories that exist worldwide. Solutions must be found that protect and promote Indigenous rights and relationships with data at scale.

Second, repositories recognize that their role in the application of CARE is limited to particular parts of the data lifecycle ([Figure 1](#)). Repositories steward data and create enabling environments for

preservation and sharing in a responsible manner. Should they be caught between competing interests, for example where an Indigenous community would like to restrict access to openly accessible third-party data (e.g., a government agency, a scientific project), it is not necessary for the repository to be the arbiter of the dispute. However, the repository may put the parties in contact with each other so that they can decide on an acceptable, actionable solution, while ensuring data preservation in the interim. In an instance where an Indigenous community feels that the use of data from a repository has been inappropriate and caused harm, there may be little a repository can do beyond recording the communication, directing the community towards the user and submitter of the data, and providing a basic assurance that the data were accessed in a manner consistent with the legal and ethical provisions associated with it.

Third, the CARE Principles apply to the entire data lifecycle, but repositories are rarely involved in all stages ([Figure 1](#)). This fundamental mismatch makes it difficult for many repositories to ensure that all of their data align with CARE. To increase the likelihood of CARE-alignment, data management planning should begin early, the CARE Principles should be integrated into data management plan (DMP) templates, and DMPs also should cover plans for data archive in an appropriate repository. If funding agencies require and enforce collaboration with Indigenous Communities prior to funding research some of the burden of tracking and verifying researcher compliance could be simplified (e.g. NSF PAPPG ([NSF 2024](#))). However, this currently is only feasible for data about a community as opposed to data of potential interest to many communities, unless all communities are in agreement.

Fourth, the subject of metadata standards here is complex, especially if we are to avoid imposing ‘cognitive imperialism’ ([Montenegro 2019](#)). Indigenous data governance and standards organizations need to be established so that they can provide guidance to the Earth Science Repository community in implementing metadata standards that address the needs of Indigenous communities. For effective implementation, these organizations will need to create well defined specifications for all of the metadata fields, semantics, and rules implied – for example, by the recently published Indigenous Metadata Bundle ([Taitingfong et al. 2023](#)) as well as the content of any TK or BC Labels. Some existing metadata standards could be promising points of departure e.g., PREMIS ([Library of Congress 2023](#)) or ODRL ([W3C 2018](#)) but it is not yet clear whether they are fully adaptable. Clarifying data’s particular cultural setting, plus potential reciprocity and obligations incurred when accessing it may be complex, depending on the granularity of metadata elements ([Christen 2012](#)) and the intended meaning imposed by the affected Indigenous communities. Furthermore, concepts of ownership, property rights, copyright law as it relates to misappropriation and metadata categories may conflict with TK and its associated religious, spiritual, or worldview meanings ([Scassa and Taylor 2017](#)).

Fifth, throughout, we have stressed the importance of working with Indigenous Peoples in any implementation. This sort of engagement has already overloaded certain individuals and groups like the Collaboratory, and will continue to do so. Communication strategies that defend against this possibility and also provide for diverse representation of Native voices would be wise; for example, creating representative bodies to communicate with multiple repositories and multiple communities would improve efficiency. While repository-to-community examples are rare, it should be noted that the Canadian territories have been working on the issue of communications with researchers for years (see ‘Conducting Research in Canada’s North’ ([Canada 2018](#))) and Tribal Nations in the US have their own laws, policies, and practices for research oversight, as well ([Garba et al. 2023; Carroll et al. 2022](#)).

Lastly, we recognize that most of the activities listed here cannot be covered by repositories within their current budgets. Many repositories have little or no budget for outreach, nor for communication beyond technical support ([USGAO 2022](#)). When they do, it is highly unlikely to cover working with all the Indigenous communities that a repository’s holdings may potentially be connected to. Foundations, agencies, universities, and other institutions that fund the creation and maintenance of repositories must appropriately support the transformation of repositories to implement CARE.

WHERE SHOULD REPOSITORIES START?

Although repositories may desire to implement sophisticated documentation of the CARE Principles, a limited workforce or limited funding and shortfalls in metadata specifications

may make it difficult to do so in a timely manner. A goal is for a repository to apply the CARE Principles to their data service and management in a manner that respects their development and maintenance resources.

The first course of action should be to determine if you already have Indigenous data in your repository. A simple first step would be to determine if the geographic coverage of existing datasets overlaps with culturally relevant or known Indigenous areas, for instance by using resources such as Native Lands Digital (<https://native-land.ca/>). Parallel (or later) steps could involve making public disclosures about the presence of Indigenous data within your repository with tools such as the Local Contexts 'Open to Collaborate' Notice, or engaging in targeted outreach to communities who are now actively asserting access and governance in their ancestral homelands, particularly around data and plant specimens (Mc Cartney et al 2022). Similarly, identifying already-deposited datasets with related collection permits or other documentation and re-curating them can help the repository community plan mechanisms for linking these resources widely. It's likely that most repositories hold datasets that meet these criteria, but have not yet characterized their holdings this way.

The next steps may depend on the type of data held. If your data are global satellite remote sensing data, your activities will be very different than for a repository holding data or collected from a particular Indigenous community's land. In the former case, the next step would be to address the technical and protocol elements to ensure the repositories' responsiveness to both Indigenous data and information requests from Indigenous communities. Following that, active engagement could determine if any global groups with both Indigenous and repository representatives have developed a set of general guidelines for this situation (at the time of this writing, the answer was "no"). Without that in place, the implementation of CARE within your repository is likely to remain a work in progress.

In the latter, an assessment of your repository's existing technical capabilities should help ascertain whether you already have systems that can handle all of the technical activities described above. For example, do you already have the ability to embargo data, apply TK Labels, attach documents and other information to data, etc.? What about your communication strategies and capabilities? Is there coordination among linked repositories or data aggregators to handle the transfer of new metadata? Do you have an advisory board and if so what communities are represented? Do you have funding for engaging with new communities? Do you have all of the needed policies and the procedures that implement them or do these need to be developed or updated to be more responsive to the CARE Principles (see *Repository Protocols* above)? In all likelihood you will need to prioritize the work needed to address any gaps, though we advise against investing in expensive upgrades until a community asks for them.

FUTURE WORK

The recommendations described here are a starting point for repositories focused on Earth and environmental science data. It is likely that over the next few years additional, more comprehensive guidelines will be developed for specific kinds of repositories. From the examples above it is clear that implementing CARE will be quite different for repositories that deal with large global datasets versus those that deal with site-specific data. It is also likely that there will need to be domain- or region-specific guidelines.

Dealing with these questions of scale is another area where future work is required. For example, one solution might be for groups or consortia of Indigenous Peoples to work with groups such as ESIP that have representatives from many repositories. In this way, rather than a series of one-on-one conversations, broadly adopted agreements amongst groups of repositories and communities could be implemented.

Another possibility would be to encourage repositories to make their data holdings widely discoverable through federated (meta)data systems. If their metadata included accurate spatial and topical metadata (including, where relevant, TK Labels and similar markup), other systems and communities could become aware of their content. That would allow communities to potentially notify the repository that it may be holding Indigenous data, as the community runs across them. This approach would support *situational awareness* through both community building and technical means, while allowing for potentially restricted access.

Similarly, if there were mechanisms to track regions, data, or subjects of general interest to Indigenous communities, repositories could use this to help identify their holdings that require application of the CARE principles. The Ocean Data and Information System (ODIS), coordinated by the International Oceanographic Data and Information Exchange (IODE) of UNESCO's Intergovernmental Oceanographic Commission (IOC), is partnering with Local Contexts to prototype just such an approach, leveraging standard web architecture and open standards. The aim is to allow all ODIS partners (and any other system using compatible approaches) to detect that they may have data in need of CARE evaluation, based on the input of Indigenous communities and data systems. Alternatively, mechanisms for users to flag datasets as needing CARE application could also be explored.

The CARE Principles and developing practices discussed here could serve as a model for other groups of stakeholders who share similar concerns but lack the political standing of Indigenous Peoples. As a framework, CARE is particularly powerful for use by marginalized populations as it highlights the importance of control and trust (Carroll et al. 2020). Further, some of the technological features described here to enable CARE are also listed in general guidelines for repositories (CoreTrustSeal Standards and Certification Board 2022; Lin et al. 2020). Therefore repository funders should view these investments in infrastructure as triply beneficial: not only are these basic technical upgrades that help enable CARE-related ethics, they also position repositories to meet other more general repository guidelines, and are likely to be transferable to engagement with other previously underrepresented communities.

CONCLUSION

Applying the CARE Principles to repository activities is a complicated process, and we have not covered all eventualities here. The overriding feature of these recommendations is the responsibility of data stewards to ensure that their repositories are responsive to Indigenous data, and the importance of communication with Indigenous Peoples when and as they want to engage, to build the deeper relationships that must underlie a repository's assertion of conforming to CARE Principles. Repositories will have to consider these recommendations in light of their data holdings, the Indigenous Peoples to whom the data relate, and the repository operating framework. The work presented in this paper is meant to be only a starting point towards building relational accountability.

DATA ACCESSIBILITY STATEMENT

Data to support [Figures 3–6](#) is available at O'Brien, M., R. Duerr, R. Taitingfong, A. Martinez, L. Vera, L. Jennings, R. Downs, et al. 2024. "Alignment between CARE Principles and Data Repository Activities." Environmental Data Initiative. <https://doi.org/10.6073/PASTA/23E699AD00F74A178031904129E78E93>.

Examples of linked activities in a project management context are available in Duerr, R. 2024. "Examples of CARE-Related Activities Carried out by Repositories, in Sequences or Groups." Environmental Data Initiative. <https://doi.org/10.6073/PASTA/1B812B3BD296D23C4C7C54EBO22774FC>.

SUPPLEMENTAL MATERIAL

[Figures 3–6](#) are presented together in tabloid form and as a list at: Duerr, R., M. O'Brien, R. Taitingfong, A. Martinez, V. Lourdes, L. Jennings, R. Downs, et al. 2024. "Earth and Environmental Science Repository Activities in Support of the CARE Principles," January. <https://zenodo.org/records/10521041>.

ETHICS AND CONSENT

SRC has given consent for use of [Figure 2](#), which originally appeared in Carroll, S.R., I. Garba, O.L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, et al. 2020. "The CARE Principles for Indigenous Data Governance" 19 (1): 43. <https://doi.org/10.5334/dsj-2020-043>.

ACKNOWLEDGEMENTS

At least 60 individuals attended online meetings, contributed ideas during presentations, or were early readers of this paper. These include several additional members of the Collaboratory for Indigenous Data Governance and the ESIP Sustainable Data Management cluster. We would like to especially thank Shelley Stall for starting us on this path and Philip Tarrant who contributed many ideas early in the project.

O'Brien et al.
Data Science Journal
DOI: 10.5334/dsj-2024-037

24

DISCLAIMER

The views expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of any part of the U.S. Government.

FUNDING INFORMATION

RRD is supported by NASA contract 80GSFC23CA001 and NASA Cooperative Agreement 80 NSSC23M0179. MO is supported by US National Science Foundation (NSF) grant numbers 1931174, 1831937, NASA grant NNX14 AR62 A and BOEM Agreement MC15 AC00006. SRC, AM are supported by the Morris K. Udall and Stewart L. Udall Foundation. RT is supported by the Henry Luce Foundation. PLB is supported by the European Union's Horizon Europe Programmes, via grants: WorldFAIR (GA:10105839), MARCO-BOLO (101082021), and BiOceans5D (101059915).

This work is based on materials, programs, collaboration platforms and meeting space provided by the ESIP Community with support from the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA) and the United States Geological Survey (USGS).

COMPETING INTERESTS

Robert Downs is a member of the Editorial Board of the Data Science Journal.

AUTHOR CONTRIBUTIONS

Conceptualization – Shelley Stall, Margaret O'Brien, Robert R. Downs, Ruth Duerr

Data curation – Margaret O'Brien

Formal analysis – Robert R. Downs, Ruth Duerr, Margaret O'Brien, Lydia Jennings, Stephanie Russo Carroll, Andrew Martinez, Riley Taitingfong

Investigation – Ruth Duerr, Margaret O'Brien, Robert R. Downs, Erin Antognoli

Methodology – Robert R. Downs, Margaret O'Brien, Ruth Duerr

Resources – Lourdes, Talya ten Brink

Software –

Validation – Lydia Jennings, Andrew Martinez, Dominique David Chavez, Stephanie Russo Carroll, Lourdes, Talya ten Brink, Nicole B. Halmai

Visualization – Ruth Duerr, Lourdes, Lydia Jennings, Robert R. Downs

Writing – original draft – Lourdes, Margaret O'Brien, Ruth Duerr, Stephanie Russo Carroll, Erin Antognoli, Robert R. Downs, Talya ten Brink

Writing – review & editing – Margaret O'Brien, Ruth Duerr, Robert R. Downs, Erin Antognoli, Talya ten Brink, Lourdes, Lydia Jennings, Riley Taitingfong, Andrew Martinez, Stephanie Russo, Carroll, Nicole B. Halmai, Pier Luigi Buttigieg, Maui Hudson, Dominique David-Chavez

AUTHOR INFORMATIONS

MO is a non-Indigenous scientific data manager and data specialist from Southern California. She works with the US Long-Term Ecological Research (LTER) Network, US Marine Biodiversity Observation (MBON) Network, and the Repository of the Environmental Data Initiative.

RD is a non-Indigenous researcher/data manager with a wide range of international experience including Indigenous data management with the Exchange for Local Observations and Knowledge of the Arctic, membership on the Arctic Data Committee, and leadership roles in both ESIP and the Research Data Alliance.

O'Brien et al.
Data Science Journal
DOI: 10.5334/dsj-2024-037

25

RT is a Chamoru (Indigenous peoples of the Mariana Islands) woman with ties to the island of Guåhan/Guam. She is also of mixed European descent and lives in San Diego on Kumeyaay lands. She is a postdoctoral scholar at the Native Nations Institute, where her work focuses on building tools and guidance to implement the CARE Principles in data repositories.

AM is a citizen of the Salt River Pima-Maricopa Indian Community. He is also Yaqui and Diegueno/lipay with family from the Mesa Grande Band of Mission Indians in San Diego County, California. He serves as the Research Coordinator with the Collaboratory for Indigenous Data Governance.

LV has no tribal membership and is of Latinx, including Kichwa, Taíno, and African, and Western European descent. She is a researcher and educator living on Haudenosaunee land, serving on the Advisory Committee of the Environmental Data and Governance Initiative.

LLJ is a citizen of the Pascua Yaqui Tribe (Yoeme) Huichol (Wixárika) and Greek. She is an environmental soil scientist at Arizona State University (Akimel O'odham and Pee Posh lands) and Research Fellow at Duke University (Tutelo, Saponi, Haliwa-Saponi, Sappony and Occaneechi lands), with affiliation to the Collaboratory for Indigenous Data Governance (O'odham and Yaqui lands).

RRD has no tribal affiliation, studies and manages development, evaluation, sustainability, use, and impact of data repositories, and co-authored the TRUST Principles for Digital Repositories.

EA has no tribal affiliation. She is a data management and metadata librarian with extensive experience working with scientific data repositories and researchers in the U.S. Department of Agriculture Agricultural Research Service, her employer during the bulk of this project. Erin Antognoli is currently employed as a data specialist with the Congressional Research Service (CRS). This article was prepared prior to her CRS employment. The views expressed herein do not represent those of CRS or the Library of Congress.

TtB is a non-Indigenous geospatial scientist, social scientist, and regional data governance coordinator at NOAA Fisheries, and chairs the International Council for the Exploration of the Sea's Working Group on Marine Spatial Planning and Coastal Zone Management. She is part of the Greater Atlantic NOAA Fisheries Indigenous Engagement, Partnerships and Knowledge (IEPK) Community of Practice and the NOAA-wide IEPK Community of Practice. She is a research associate of the University of Rhode Island. The views expressed herein are those of the authors and do not reflect the official policy or position of National Oceanic and Atmospheric Administration (NOAA), Department of Commerce or the U.S. Government.

NBH is a citizen of the Navajo Nation with Navajo, Choctaw and mixed European ancestry. She is a cancer health disparities researcher and the lead for the Native and Indigenous cancer health task force within the Center for Advancing Cancer Health Equity at the UC Davis Comprehensive Cancer Center.

DDC is multicultural Indigenous Caribbean (Arawak Taíno), carrying Indigenous Boricua, African, Spanish and East European ancestry. She currently directs the Indigenous Land & Data Stewards Lab on homelands of the Arapaho, Cheyenne, and Ute Nations in the Colorado Front Range.

SRC is Ahtna, a citizen of the Native Village of Kluti-Kaah in Alaska, USA, and of Sicilian descent. She directs the Collaboratory for Indigenous Data Governance, is a co-author of the CARE Principles, and leads the US Indigenous Data Sovereignty Network and the Global Indigenous Data Alliance.

MH is a member of the Whakatōhea Nation in Aotearoa New Zealand. He is a co-author of the CARE Principles and member of Te Mana Raraunga Māori Data Sovereignty Network and the Global Indigenous Data Alliance.

PLB has no tribal affiliation. He is a digital architect, data scientist, and digital strategist specializing in creating federated and diverse digital ecosystems. He is currently working with Local Contexts in implementing the CARE Principles through his work with the IOC-UNESCO Ocean Data and Information System (ODIS).

AUTHOR AFFILIATIONS

O'Brien et al.
Data Science Journal
DOI: 10.5334/dsj-2024-037

Margaret O'Brien  orcid.org/0000-0002-1693-8322

University of California, Santa Barbara, CA, USA

Ruth Duerr  orcid.org/0000-0003-4808-4736

Ronin Institute for Independent Scholarship, US

Riley Taitingfong  orcid.org/0000-0003-0981-4783

Native Nations Institute, University of Arizona, US

Andrew Martinez  orcid.org/0000-0002-4500-4422

Native Nations Institute, University of Arizona, US

Lourdes Vera  orcid.org/0000-0002-5390-6397

University at Buffalo, The Environmental Data and Governance Initiative, US

Lydia L. Jennings  orcid.org/0000-0001-9860-5574

Lands of the O'odham and Yaqui peoples, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA; Lands of the Akimel O'odham and Pee Posh Peoples, School of Sustainability, Arizona State University, Phoenix, AZ, USA

Robert R. Downs  orcid.org/0000-0002-8595-5134

CIESIN, Columbia Climate School, Columbia University, Palisades, NY, USA

Erin Antognoli  orcid.org/0000-0003-0569-0808

Library of Congress, USA

Talya ten Brink  orcid.org/0000-0002-3279-0575

University of Rhode Island/NOAA Fisheries, USA

Nicole B. Halmai  orcid.org/0000-0002-0790-6548

University of California, Davis, Genome Center, Davis, CA, USA; University of California, Davis, Comprehensive Cancer Center, Sacramento, CA, USA

Dominique David-Chavez  orcid.org/0000-0001-5393-4331

Colorado State University, Department of Forest and Rangeland Stewardship, US

Stephanie Russo Carroll  orcid.org/0000-0002-8996-8071

Lands of the O'odham and Yaqui peoples, Mel and Enid Zuckerman College of Public Health, Native Nations Institute, and Udall Center for Studies in Public Policy, University of Arizona, USA

Maui Hudson  orcid.org/0000-0003-3880-4015

Lands of the Waikato-Tainui People, Te Kotahi Research Institute, University of Waikato, Hamilton, New Zealand

Pier Luigi Buttigieg  orcid.org/0000-0002-4366-3088

Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

REFERENCES

Anderson, A, Huttenlocher, D, Kleinberg, J and Leskovec, J 2013 Steering User Behavior with Badges. In *Proceedings of the 22nd International Conference on World Wide Web*, 95–106. WWW '13. New York, NY, USA: Association for Computing Machinery. DOI: <https://doi.org/10.1145/2488388.2488398>

Arbour, L and Cook, D 2006 DNA on Loan: Issues to Consider When Carrying Out Genetic Research with Aboriginal Families and Communities. *Community Genetics*, 9(3): 153–60. DOI: <https://doi.org/10.1159/000092651>

Avila, R 2023 Updating the Open Definition to Meet the Challenges of Today. *Open Knowledge blog* March 16, 2023. Available at: <https://blog.okfn.org/2023/03/16/updating-the-open-definition-to-meet-the-challenges-of-today/>.

Bahim, C, Casorrán-Amilburu, C, Dekkers, M, Herczog, E, Loozen, N, Repanas, K, Russell, K and Stall, S 2020 The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal*, 19: 41. DOI: <https://doi.org/10.5334/dsj-2020-041>

Baker, K S and Duerr, R E 2017 Research and the Changing Nature of Data Repositories. In Johnston, L R (ed.) *Curating Research Data*. Chicago: American Library Associationpp. 33ff.

Baker, K S, Duerr, R E and Parsons, M A 2015 Scientific Knowledge Mobilization: Co-Evolution of Data Products and Designated Communities. *International Journal of Digital Curation*, 10(2): 110–35. DOI: <https://doi.org/10.2218/ijdc.v10i2.346>

Ball, A 2012 *Review of Data Management Lifecycle Models*. Bath: University of Bath.

'BC Provenance – Local Contexts.' n.d. Accessed October 31, 2023. Available at: <https://localcontexts.org/label/bc-provenance/>.

BCO-DMO (n.d) Introduction to BCO-DMO | BCO-DMO. Accessed December 29, 2023. Available at: <https://www.bco-dmo.org/>

Belarde-Lewis, M, Littletree, S, Braine, I R, Srader, K, Guerrero, N and Palmer, C L 2024 Centering Relationality and CARE for Stewardship of Indigenous Research Data. *Data Science Journal*, 23: 32. DOI: <https://doi.org/10.5334/dsj-2024-032>

Canada, Polar Knowledge 2018 Conducting Research in Canada's North. December 6, 2018. <https://www.canada.ca/en/polar-knowledge/online-portal-for-researchers.html>.

Carlson, D J 2012 IPY 2007–2008: Where Threads of the Double Helix and Sputnik Intertwine. In Huettman, F (ed.), *Protection of the Three Poles*. Tokyo: Springer Japan, 35–50. DOI: https://doi.org/10.1007/978-4-431-54006-9_2

Carroll, S R, Garba, I, Figueroa-Rodríguez, O L, Holbrook, J, Lovett, R, Materechera, S, Parsons, M, et al 2020 *The CARE Principles for Indigenous Data Governance*, 19(1): 43. DOI: <https://doi.org/10.5334/dsj-2020-043>

Carroll, S R, Garba, I, Plevel, R, Hiratsuka, V Y, Hudson, M and Garrison, N A 2022 Using Indigenous Standards to Implement the CARE Principles: Setting Expectations through Tribal Research Codes. *Frontiers in Genetics*, 13. DOI: <https://doi.org/10.3389/fgene.2022.823309>

CCSDS 650.0-B-1 2012 REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS). CCSDS. Available at: <https://public.ccsds.org/pubs/650x0m2.pdf>.

Chapman, A 2020 Current Best Practices for Generalizing Sensitive Species Occurrence Data. *GBIF Secretariat*. DOI: <https://doi.org/10.15468/doc-5jp4-5g10>

Christen, K A 2012 Does Information Really Want to Be Free? Indigenous Knowledge Systems and the Question of Openness. *International Journal of Communication*, 6: 24ff.

CIESIN, Columbia University 2024 Socioeconomic Data and Applications Center. *Socioeconomic Data and Applications Center*. February 2, 2024. Available at: <https://sedac.ciesin.columbia.edu/>.

Collaboratory for Indigenous Data Governance n.d. Collaboratory for Indigenous Data Governance. Accessed December 11, 2023. Available at: <https://indigenousdatalab.org/>.

CoreTrustSeal Standards and Certification Board 2022 CoreTrustSeal Requirements 2023–2025. DOI: <https://doi.org/10.5281/zenodo.7051012>

Cultural Survival, and First Peoples Worldwide 2023 Securing Indigenous Peoples Right to Self-Determination: A GuIdE On Free, Prior And Informed Consent. *Cultural Survival blog*. Available at: <https://www.culturalsurvival.org/news/new-guide-supports-indigenous-leaders-develop-fpic-protocols-and-secure-self-determined>.

DataONE n.d. *Data Observation Network for Earth | DataONE*. Accessed February 9, 2024. Available at: <https://www.dataone.org/>.

Davis, M 2016 Data and the United Nations Declaration on the Rights of Indigenous Peoples. In Kukutai, T and Taylor, J (eds.), *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University (ANU) Press. DOI: <https://doi.org/10.22459/CAEPR38.11.2016.02>

Downs, R R 2021 Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories. *Data Science Journal*, 20: 1. DOI: <https://doi.org/10.5334/dsj-2021-001>

Drabiak-Syed, K 2010 Lessons from Havasupai Tribe v. Arizona State University Board of Regents: Recognizing Group, Cultural, and Dignity Harms as Legitimate Risks Warranting Integration into Research Practice. *J. Health & Biomedical L.* 6, pp. 175.

Duerr, R 2024 Examples of CARE-Related Activities Carried out by Repositories, in Sequences or Groups. *Environmental Data Initiative*. DOI: <https://doi.org/10.6073/PASTA/AEADD84525AE922D3115FD05CF5BF4FE>

Duerr, R, O'Brien, M, Taitingfong, R, Martinez, A, Lourdes, V, Jennings, L, Downs, R, et al. 2024 *Earth and Environmental Science Repository Activities in Support of the CARE Principles*. <https://zenodo.org/records/10521041>.

EDI (n.d.) *Environmental Data Initiative*. Accessed December 29, 2023. Available at: <https://edirepository.org/>.

Eitzel, M V, Cappadonna, J, L, Santos-Lang, C, Duerr, R E, Virapongse, A, West, S E, Kyba, C, Bowser, A, Cooper, C B and Sforzi, A 2017 Citizen Science Terminology Matters: Exploring Key Terms. *Citizen Science: Theory and Practice*, 2(1). DOI: <https://doi.org/10.5334/cstp.96>

ELOKA 2023 *Welcome | ELOKA*. Available at: <https://eloka-arctic.org/>.

ESIPFed (n.d.) Home – Earth Science Information Partners (ESIP). *ESIPFed*. Accessed December 28, 2023. Available at: <https://www.esipfed.org/>.

Foundation of Digital Badges and Micro-Credentials. n.d. Accessed June 3, 2024. Available at: <https://link.springer.com/book/10.1007/978-3-319-15425-1>.

Garba, I, Sterling, R, Plevel, R, Carson, W, Cordova-Marks, F M, Cummins, J, Curley, C, et al. 2023 Indigenous Peoples and Research: Self-Determination in Research Governance. *Frontiers in Research Metrics and Analytics*, 8. DOI: <https://doi.org/10.3389/frma.2023.1272318>

GIDA 2023a CARE Principles. *Global Indigenous Data Alliance*. January 23, 2023. Available at: <https://www.gida-global.org/care>.

GIDA 2023b IEEE Provenance. Global Indigenous Data Alliance. January 23, 2023. Available at: <https://www.gida-global.org/ieee-provenance>.

Gilbert, T, Williams, P, Fields, M, Williams, L, Flitt, W, Savage, M, John, C, et al. 2017 *Gwich'in Na'iin Googindi' K'it/Gwich'in Place Names Atlas* *Gwich'in Place Names Atlas*. Boulder, Colorado USA: National Snow and Ice Data Center. Available at: <https://eloka-arctic.org/product/atlas/gwichin-place-names-atlas>.

Global Indigenous Data Alliance 2023 *Global Indigenous Data Alliance*. Available at: <https://www.gida-global.org>.

Hernandez, J 2022 Fresh Banana Leaves: Healing Indigenous Landscapes Through Indigenous Science. Available at: <https://bookshop.org/p/books/fresh-banana-leaves-healing-indigenous-landscapes-through-indigenous-science-jessica-hernandez/16859402>.

Hudson, M, Carroll, S R, Anderson, J, Blackwater, D, Cordova-Marks, F M, Cummins, J, David-Chavez, D, et al. 2023 Indigenous Peoples Rights in Data: A Contribution toward Indigenous Research Sovereignty, *Frontiers in Research Metrics and Analytics*, 8: 1173805. DOI: <https://doi.org/10.3389/frma.2023.1173805>

ICPSR 2024 *ICPSR Data Excellence Research Impact*. Available at: <https://www.icpsr.umich.edu/web/pages/>.

ITK and NRI 2007 Nickels, S, Shirley, J and Laidler, G (ed.), *Negotiating Research Relationships with Inuit Communities: A Guide for Researchers*. Ottawa and Iqualuit: Inuit Tapiriit Kanatami and Nunavut Research Institute. Available at: <https://repository.oceanbestpractices.org/handle/11329/1605?show=full>

ITTAQ 2024 Clyde River Weather, *Kangiqtugaapik (Clyde River) Weather Station Network*. Available at: <https://clyderiverweather.org/>.

Jennings, L, Anderson, T, Martinez, A, Sterling, R, David Chavez, D, Garba, I, Hudson, M, Garrison, N A and Carroll, S R 2023 Applying the 'CARE Principles for Indigenous Data Governance' to Ecology and Biodiversity Research. *Nature Ecology & Evolution*, 7. DOI: <https://doi.org/10.1038/s41559-023-02161-2>

Jones, M, Slaughter, P and Habermann, T. 2019 Quantifying FAIR: automated metadata improvement and guidance in the DataONE repository network. DOI: <https://doi.org/10.5281/ZENODO.3408465>

Kanao, M, Okada, M and Kadokura, A 2014 Metadata Management at the Polar Data Centre of the National Institute of Polar Research, Japan. *Data Science Journal*, 13. DOI: <https://doi.org/10.2481/dsj.IFPDA-05>

Kukutai, T 2023 Indigenous Data Sovereignty—A New Take on an Old Theme. *Science*, 382(6674): eadl4664. DOI: <https://doi.org/10.1126/science.adl4664>

Kukutai, T and Taylor, J (eds.) 2016 *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: ANU Press. DOI: <https://doi.org/10.22459/CAEPR38.11.2016>

Library of Congress 2023 PREMIS Data Dictionary for Preservation Metadata. Available at: <https://www.loc.gov/standards/premis/>.

Lin, D, Crabtree, J, Dillo, I, Downs, R R, Edmunds, R, Giaretta, D and De Giusti, M, et al. 2020 The TRUST Principles for Digital Repositories. *Scientific Data*, 7(1): 144. DOI: <https://doi.org/10.1038/s41597-020-0486-7>

Local Contexts 2024 *Local Contexts – Grounding Indigenous Rights*. Available at: <https://localcontexts.org/>.

Local Contexts (n.d.) *TK Labels – Local Contexts*. Accessed December 29, 2023. Available at: <https://localcontexts.org/labels/traditional-knowledge-labels/>.

LTER (n.d.); Home - LTER. Accessed February 9, 2024. Available at: <https://lternet.edu/>.

MacInnes, A, Colchester, M and Whitmore, A 2017 Free, Prior and Informed Consent: How to Rectify the Devastating Consequences of Harmful Mining for Indigenous Peoples. *Perspectives in Ecology and Conservation*, 15(3): 152–60. DOI: <https://doi.org/10.1016/j.pecon.2017.05.007>

Mc Cartney, A M, Anderson, J, Liggins, L, Hudson, M L, Anderson, M Z, TeAika, B, Geary, J, Cook-Deegan, R, et al. 2022 Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. *Proceedings of the National Academy of Sciences*, 119(4): e2115860119. DOI: <https://doi.org/10.1073/pnas.2115860119>

Montenegro, M 2019 Subverting the Universality of Metadata Standards: The TK Labels as a Tool to Promote Indigenous Data Sovereignty. *Journal of Documentation*, 75(4): 731–49. DOI: <https://doi.org/10.1108/JD-08-2018-0124>

Noorbergen, T J, Adam, M T P, Roxburgh, M and Teubner, T 2021 Co-Design in mHealth Systems Development: Insights From a Systematic Literature Review. *AIS Transactions on Human-Computer Interaction*, 13(2): 175–205. DOI: <https://doi.org/10.17705/1thci.00147>

NSF 2024 *Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 24-1) | NSF – National Science Foundation*. Available at: <https://new.nsf.gov/policies/pappg/24-1>.

NSIDC 2023 *Homepage*. *National Snow and Ice Data Center*. Available at: <https://nsidc.org/home>.

O'Brien, M, Duerr, R, Taitingfong, R, Martinez, A, Vera, L, Jennings, L, Downs, R, et al. 2024 Alignment between CARE Principles and Data Repository Activities. *Environmental Data Initiative*. DOI: <https://doi.org/10.6073/PASTA/23E699AD00F74A178031904129E78E93>

OKFN 2016 Open Data Handbook. *Open Knowledge Foundation*. Available at: <https://opendatahandbook.org/guide/en/>.

RDA 2017 *International Indigenous Data Sovereignty IG*. Available at: <https://www.rd-alliance.org/groups/international-indigenous-data-sovereignty-ig>.

Royal Society Te Aparangi 2023 *Mana Raraunga Data Sovereignty*. Available at: <https://www.royalsociety.org.nz/assets/Mana-Raraunga-Data-Sovereignty-web-V1.pdf>.

Scassa, T and **Taylor, F** 2017 Legal and Ethical Issues around Incorporating Traditional Knowledge in Polar Data Infrastructures. *DataScienceJournal*, 16: 3, 1–14. DOI: <https://doi.org/10.5334/dsj-2017-003>

Shingledecker, S, Mills, A, Orlando, M, Burgess, A, Allen, P and **Cruz, E** 2024 2023 Annual Report for Earth Science Information Partners (ESIP). *ESIP*. DOI: <https://doi.org/10.6084/m9.figshare.25114319.v1>

Smithsonian Institution 2019 Smithsonian Directive 609 – Digital Asset Access and Use. *Smithsonian Institution*. Available at: <https://www.si.edu/content/pdf/about/sd/SD609.pdf>.

Taitingfong, R, Martinez, A, Russo Carroll, S, Hudson, M and **Anderson, J** 2023 Indigenous Metadata Bundle Communiqué. *Collaboratory for Indigenous Data Governance, ENRICH: Equity for Indigenous Research and Innovation Coordinating Hub, and Tikanga in Technology*. Available at: <https://docs.google.com/document/d/1XJReoSqazzp8DI2u2NLnnwlhoJvXEI0BqAmt0wupU4/edit>.

TK Labels – Local Contexts. n.d. Accessed October 31, 2023. Available at: <https://localcontexts.org/labels/traditional-knowledge-labels/>.

United Nations n.d.-a. Indigenous Peoples. *United Nations*. Accessed January 11, 2024. Available at: <https://www.un.org/en/fight-racism/vulnerable-groups/indigenous-peoples>.

United Nations n.d.-b. *United Nations Declaration On The Rights Of Indigenous Peoples*. Accessed March 26, 2024. Available at: <https://social.desa.un.org/issues/indigenous-peoples/united-nations-declaration-on-the-rights-of-indigenous-peoples>.

USGAO 2022 *Tribal Funding: Actions Needed to Improve Information on Federal Funds That Benefit Native Americans* | U.S. GAO. Available at: <https://www.gao.gov/products/gao-22-104602>.

Vierkant, P (n.d.) DataCite – Connecting Research, Advancing Knowledge. *DataCite*. Accessed February 9, 2024. Available at: <https://datacite.org/>.

W3C 2018 *ODRL Information Model 2.2*. Available at: <https://www.w3.org/TR/odrl-model/>.

Webb, R, Bai, X, Stafford Smith, M, Costanza, R, Griggs, D, Moglia, M, Neuman, M, et al. 2018 Sustainable Urban Systems: Co-Design and Framing for Transformation, *Ambio*, 47(1): 57–77. DOI: <https://doi.org/10.1007/s13280-017-0934-6>

Wilkinson, M D, Dumontier, M, Aalbersberg, I J, Appleton, G, Axton, M, Baak, A, Blomberg, N, et al. 2016 The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data*, 3(1): 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>

Youning, G 2018 *Elements of Indigenous Style: A Guide for Writing by and about Indigenous Peoples*. Indigenous Collection. Edmonton, Alberta: Brush Education.

O'Brien et al.
Data Science Journal
DOI: 10.5334/dsj-2024-037

29

TO CITE THIS ARTICLE:

O'Brien, M, Duerr, R, Taitingfong, R, Martinez, A, Vera, L, Jennings, L L, Downs, R R, Antognoli, E, ten Brink, T, Halmaj, N B, Carroll, S R, David-Chavez, D, Hudson, M, and Buttigieg, P L 2024 Earth Science Data Repositories: Implementing the CARE Principles. *Data Science Journal*, 23: 37, pp. 1–29. DOI: <https://doi.org/10.5334/dsj-2024-037>

Submitted: 29 March 2024

Accepted: 17 June 2024

Published: 03 July 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.