Polar Research Data Management: Understanding Technical Implementation and Policy Decisions in the Era of FAIR Data

Gregory Vey, Wesley Van Wychen, Chantelle Verhey, Peter Pulsifer, Ellsworth LeDrew

From *Library and Information Sciences in Arctic and Northern Studies*, Springer Polar Sciences (In Press).
Corresponding Author: gvey@uwaterloo.ca

Metadata Standards

The polar research data management (RDM) community operates in a multidisciplinary space. Options for standards are numerous, detailed and cover information about the data itself, the machines/instruments used to collect the data (such as make, model, and manufacturer), as well as any cleaning and/or analysis steps or scripts used to process or create data products. Hand in hand with metadata standards, controlled vocabularies supply standardized sets of property values, such that common metadata elements (i.e. fields) and their respective values (i.e. contents) uniformly describe the features using the same terms. Combining both strategies improves the ability of an individual or machine to find the data for which they are searching, while also providing the opportunity for communities of practice to actively share data. All of these factors are pertinent to curating robust metadata. Due to the complexity and cross domain nature of the data, a single metadata standard has yet to emerge for the polar RDM community. There have been initiatives to create a universal metadata standard in the past. However, pursuing a "one-size-fits-all" solution has so far proven unrealistic, as adoption requires broad consensus among heterogeneous stakeholder groups. Thus, stemmed a more collaborative approach of establishing crosswalks between standards (see Semantics and Interoperability). Repositories currently have no obligation to choose a standard or 'schema' from the existing inventory, and some organizations and repositories even elect to create their own formats, such as the Socioeconomic Data and Applications Center's CEISIN standard . From the vast number of standards that have emerged, there are several that are more common throughout the polar RDM community, including International Organization for Standardization (ISO) 19115 in XML, Federal Geospatial Data Committee (FGDC) in XML, GCMD Directory Interchange Format (DIF) in XML, DataCite Schema in XML, Dublin Core in XML, and Data Catalog Vocabulary (DCAT) in JSON. Note that format (e.g. XML, JSON) is irrespective of standard (see Semantics and Interoperability).

In the case of PDC, the ISO 19115 standard is the core standard used to implement metadata, along with translations to other formats, like FGDC. The PDC uses a standardized ISO 19115 web schema for its users to fill out alongside their datasets. Once submitted a dedicated data manager reviews the records for consistency and robustness. However, this process can vary from repository to repository. Although, the ISO 19115 standard is compatible with the FGDC and DIF standards to abide by the Antarctic treaty. ISO 19115 was established by the International Standards organization in 2003, and is predominately used within the earth science domain, and geospatial data in particular. The standard allows for a more tailored metadata record. For example, ISO 19115 provides information about the description, the extent, the quality, the spatial and temporal schema, spatial reference, keywords, and distribution of digital geographic data (source). However, it should be noted that while the ISO 19115 standard is well known and well documented, it is a proprietary standard, and potential implications of this facet should be well understood by adopting repositories (see Organizational Policy).

As standards continue to evolve, PDC stays engaged with the current community trends and requirements to ensure its metadata and data holdings meet these needs. A current example is the adoption of the FAIR Data Principles. Currently, PDC is challenged by how to reconcile legacy infrastructure design decisions to best implement this paramount policy shift, as it continues to emerge for the betterment of the data management landscape. The sections that follow will elucidate the various factors and considerations raised by implementing FAIR data, from a variety of relevant perspectives.

Data Architecture

Data architecture represents a core consideration for any organization. Decisions with respect to policy and technical implementation will have a tremendous cascading impact on how organizational activities are executed and will shape how an organization interacts with its intended usage domain and target users. Furthermore, decisions on data architecture are difficult to retract or modify, both in terms of policy and technical implementation, therefore this is typically something that must be well planned in the formative stages of the organization. In the case of PDC, the implementation of geographic metadata standards, the heterogeneity of the contributed data, and repository future-proofing, all exert an effect on data modelling and data architecture perspectives, policies, and implementation.

Influenced by the factors listed above, PDC implements metadata architecture using a normalized relational model . This approach is also used for PDC data and other assets, such as the RADARSAT collections. The relational model (RM) offers benefits such as low data redundancy, data consistency, and physical data independence. In terms of implementation, the RM rests on a formal mathematical basis and relational database (RDB) design is accomplished through a formal normalization process. However, there are multiple aspects of data representation and entity-relationship (ER) modelling that are not effectively captured by the RM.

RDBs are unable to directly represent many real-world objects, especially those that are complex and composed of other objects. This is a result of the inability of the RM to distinguish between entities versus relationships, since relationships identified during ER modelling are not directly represented within the RM, and therefore do not persist in an explicit fashion. This prevents the RM from offering a means to directly recover the relationships between entities, such as the Works In relationship between Employee and Department entities. Consequently, users must possess prior knowledge about such relationships, in order to compensate for the semantic overloading that occurs, because relations from the RM are used to represent both the entities and the relationships from the corresponding ER model.

There are a variety of other challenges for the RM that include excessive fragmentation that can require numerous joins to meet queries, the inability directly capture lists or sets, and the inability to directly include a composite attribute, such as Name, which might contain member attributes like First Name and Last Name. Related to this, the range of available datatypes is limited and there is no way to create user-defined types required for specific application domains. Similarly, the RM cannot depict hierarchical or inheritance associations, such as inferring that entities like Employee and Student both inherit the attributes of a mutual parent entity like Person, or that the set of all Employees is a subset of all Persons.

In order to better appreciate the technical formalities discussed so far, we can consider a usage scenario taken from the polar RDM domain. Figure 1 shows an example of an ER diagram modeling a polar metadata record, based on the actual RM implementation used at PDC. The sample diagram shows the fields (i.e. columns) contained in the primary Metadata table in the centre, as well as three other associated tables, each with their own respective fields. In particular, note that the first field (i.e. the primary key) in the Metadata table, metadata_id, also occurs as a field in each of the three associated tables. This design provides a retrieval process where the metadata_id associated with any metadata record can be used to lookup additional information, like the Research Program associated with that particular instance of metadata.

Decoupling the data across associated tables is the result of the previously mentioned process of normalization and provides benefits such as low data redundancy, while enforcing strong data integrity. However, if you needed to generate a report on the contents of current metadata, this would necessitate fetching back information from the associated tables through numerous Join operations. For example, if the report requires the name of Research Program or the name of the Responsible Party for each record, these are not directly contained within the primary Metadata table, therefore increasing the overall retrieval cost to obtain this more comprehensive information.

Despite the previously described limitations, it is important to recognize that the RM and RDBs offer a gold standard of reliability and serve as the core data persistence and data management components of a vast number of commercial, financial, and academic organizations. Nevertheless, the growing adoption of non-relational, or Not Only SQL (NoSQL), database implementations lends motivation to consider possible advantages, particularly with respect to trends such as big data. In particular, some of the key advantages of NoSQL databases include distributed database capability, horizontal scalability, and schema-free insertion of data. Schema-free representation of data is an especially important capability for many application domains, and is of high interest to PDC data architecture considerations.

One example of an issue that has prompted reconsideration of PDC metadata architecture is the assignment of Digital Object Identifiers (DOIs). A DOI is an ISO standardized persistent identifier that is used to uniquely identify various objects, such as published datasets. In the case of PDC, DOI assignment is an organizational practice that began after the design and implementation of the metadata tables that comprise the internal representation and storage of PDC metadata. This consequently resulted in an alteration to the specification of internal metadata. In terms of implementing this change, one solution would be to alter the core metadata table by adding a column to store DOI values. However, this would introduce notable risk for a live production system, even after thorough testing in an offline version. In addition, there is an issue of sparsity as many previous records might not have values for this column. An alternative to altering the core metadata table would be to create a separate table that has only two columns: the unique metadata record identifier from the core table and the DOI associated with it. This approach both reduces risk and resolves sparsity, but it violates the concept of normality that is central to using an RDB. This is because the unique metadata record identifier is used as the primary key in both cases, and therefore, according to ER modeling, the DOI value should be a field (i.e. column) within the core metadata table, rather than a field in an ancillary table. In contrast, using a schema-free NoSQL would have allowed for the direct inclusion of DOI values within impacted metadata records, starting whenever this organizational decision was adopted.

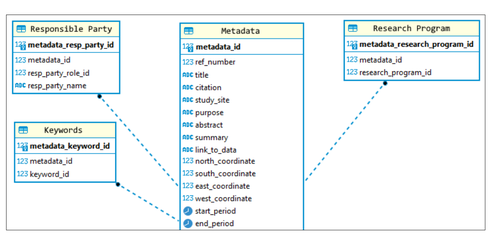

Figure 1 – Sample ER Diagram

Data Architecture (Continued)

The previous example shows how a NoSQL implementation can offer crucial flexibility in the event of a changing data model. However, there are costs and benefits associated with using a NoSQL implementation versus an RDB. RDBs are well suited for usage domains where the data are structured and low data redundancy is important. Furthermore, RDBs are implemented to support a broad range queries with good performance (although excessive joins can impact this) and this feature can be useful if new queries are likely to be required with emerging organizational needs. In contrast, NoSQL is not bound by a fixed schema and is therefore well suited for semi-structured or irregular data. However, this flexibility can result in redundancy and does not necessarily support general purpose querying. This makes a NoSQL implementation, like a MongoDB document-oriented database, a good fit when a recurrent usage domain is well understood, such as a web-based application that fetches and updates user profiles and the artifacts associated with a user account. This last perspective aligns well with the concept of maintaining PDC user profiles and their associated metadata and data contributions.

Another area of interest for PDC, with respect to MongoDB, relates to the JSON-like structure of the stored documents. JSON (JavaScript Object Notation) is a self-describing, lightweight format for storing and transporting data, and is commonly used in web and client-server application development (see Dissemination and Knowledge Mobilization). In particular, the ease of conversion between JSON used for APIs and data driven web apps versus the internal binary JSON (BSON) representation within MongoDB is also a noteworthy feature, when considering data architecture. This is discussed further in the next section on interoperability. So far, the present discussion of data architecture and technological implementations has been driven by conventional organizational pursuits. As a final thought on data architecture, we would like to draw attention to potential limitations and challenges faced when a key goal is the inclusion and representation of Indigenous Knowledge, especially within the context of Indigenous Knowledge paradigms themselves. Based on the discussion above, it seems that a schema-free design might be more successful than a structured table approach, and such designs have been previously implemented as solutions in this domain. However, design and implementation ultimately depend on the perspective of how content should be best decomposed, if at all, and represented for internal storage and subsequent reconstruction. We offer no immediate recommendation here, but we do believe that this topic embodies a crucial issue that should be central to future data architectures.

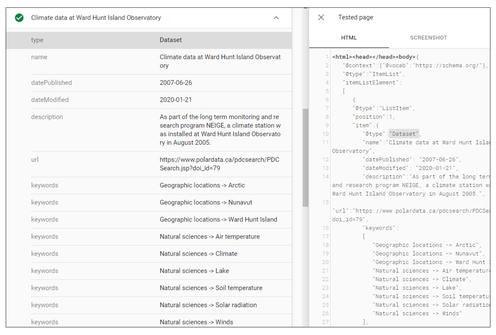

Figure 2 – PDC JSON-LD Metadata Topology

Semantics & Interoperability

Recent years have seen a strong interest in FAIR data as a premise, while also noting the disparity between advocating for policy versus actual technical implementations. This sentiment was reiterated by the European Commission Expert Group on FAIR Data report entitled Turning FAIR Into Reality which does address the well established concept of a technical ecosystem. However, while recognizing that FAIR data implementations involve technical facets, these works are not aimed at providing stepwise guidance for software development required to FAIR data compliance.

Large RDM systems and metadata platforms do exist that explicitly address FAIR data management considerations, such as the German Network for Bioinformatics Infrastructure, National Science Foundation Arctic Data Center, Ocean FAIR Data Services, and several others in areas such as the life sciences and earth sciences. However, again, the focus of these ventures relates to the provision of FAIR capabilities, rather than an exhaustive presentation of how to carry out technical implementations and/or refactoring of existing RDM systems.

Interoperability is a key component of the FAIR Data Principles. Interoperability in RDM entails systems exchanging services and data with one another. This component of the FAIR principles is essential for large scale user consumption and information exchange between repositories. Linked open data (LoD) follows a set of design principles for sharing machine-readable interlinked data on the web, and can be freely used and distributed. Although, challenges that arise from these information exchanges include systems having to agree upon common standards, algorithms, and in general, semantics. The advantage of using LoD is that a single model (Resource Description Framework, RDF) and associated formats (e.g. JSON-LD) can be used to represent metadata and data, and explicitly define the semantics (labels, definitions, hierarchical and logical relationships etc.) of the (meta)data represented.

One solution that the W3C community has utilized in order to establish a common light-weight metadata standard that can be utilized on a large-scale basis is schema.org (i.e. SDO). SDO was created in 2011 by the five major search engine organizations such as Google, Yahoo, Bing, and Yandex. Typically, SDO is utilized to mark-up web pages to enhance their discoverability in search engines such as Google. Utilizing this concept of LoD, the published datasets described by schema.org can be structured in a way that they can be linked to other datasets and resources. Google has also established a Google Dataset Search, which aims to do what Google Scholar has done for published articles, for dataset discoverability. They have provided guidance on how repositories can mark-up their existing metadata schemas in order to be included in their search interface.

The RDM community has since identified semantic mark-up as the most viable way to enhance interoperability between and among data repositories. Specific to the polar community, POLDER is working to find the best path forward for our community relating to data discovery. The international working group has been collaborating for the past five years to establish a federated search tool. POLDER is a collaboration between the Southern Ocean Observing System, Arctic Data Committee, and Standing Committee on Antarctic Data Management. Federated metadata search for the polar regions will dramatically simplify data discovery for polar scientists. Instead of searching dozens of metadata catalogues separately, a user can come to a single search page. Other domains have already completed similar projects, although challenges unique to the polar community has delayed the full adoption development. First, is the polar community being a diverse and dynamic community that produce multidisciplinary datasets and data from local communities in the Arctic, including Indigenous data and information. During the International Polar Year (2007-2009) a common metadata profile was developed and implemented by several data centres, including the PDC. PDC proceeded to implement an aggregration tool that connected many catalogues. Similarly, the Arctic Data Explorer was launched in 2013 and operated until 2019. These tools were useful , however, the development was driven by specific data centres and funding rather than a full, community-driven initiative and broad adoption was not realized. The POLDER effort builds on community-wide collaboration from the outset and builds on SDO, a model that is increasingly being used across many communities of practice. For example, since the release of the basic SDO vocabulary, the environmental science domain has created a 'science-on-schema.org' set of guidelines and recommendations, which is being utilized to guide polar repositories in order to be included in the polar federated search tool. The guidelines provide guidance for members of that domain to complete mark-up with only the relevant terms that are applicable to their schemas, no matter which one the repository is using. Second, the polar federated search tool has developed slowly due to lack of capacity and funding within the community. Specifically funding for repositories to dedicate development resources on their own infrastructure needs in order to include semantic mark-up within their existing landing pages, or for the creation of one if needed. For example, funding for Antarctic initiatives and repository management is scarce, further exacerbating the problem that is already so prevalent. In order to alleviate some of these issues, the POLDER group is working together to raise funds and provide web developmental resources on a pro bono basis to have underfunded repositories hosting valuable data, included in the federated search tooling.

Acknowledgements: The authors acknowledge financial support from the following organizations and programs: Amundsen Science, Université Laval; CFI-Cyberinfrastructure (Canadian Consortium for Arctic Data Interoperability), University of Calgary; Institutional Support (Canadian Consortium for Arctic Data Interoperability), University of Waterloo; Northern Contaminants Program/Crown-Indigenous Relations and Northern Affairs Canada; Nunavut General Monitoring Plan/Crown-Indigenous Relations and Northern Affairs Canada; Polar Knowledge Canada/Canadian High Arctic Research Station.

Dissemination & Knowledge Mobilization

Over the past few years, the single most requested new feature from PDC users has been the development of a server-side web application programming interface (API) to expose PDC metadata through a collection of endpoints. In particular, RESTful web APIs offer a means to handle stateless requests made to an endpoint through a response payload that supports a specific format, such as JSON or XML . Therefore, the development of a RESTful web API for PDC metadata is a key opportunity to meet user needs, while the implementing some of the previously discussed issues relating to interoperability. Specifically, the implementation of JSON-LD endpoints provides a mechanism to expose metadata while supporting schema.org capability and facilitating findability through search engine optimization (SEO). JSON for Linking Data (JSON-LD) is a JSON implementation of the previously mentioned RDF model that is the foundation of the LoD. Implementing equivalent XML responses and/or endpoints is also an essential consideration for PDC, in order to support legacy formats, specifically the common ISO 19115 XML standard.

Internet search engines have become a common tool in the discovery and acquisition of research datasets. As a result, SEO has become an increasingly relevant pursuit for data repositories. SEO involves optimizing web pages such that the algorithmic search results of relevant keywords will increase the rank of the given webpages. In turn this improved ranking leads to improved discoverability of assets, and in the case of data repositories, improved findability of data, from the FAIR principles perspective. These considerations of SEO have impacted the PDC metadata API implementation such that the JSON-LD endpoints provide schema.org compliant responses. This is the result of intentional congruence between the JSON representation that is designed for embedding in a webpage corresponding to a PDC metadata record and the JSON representation of the same record provided as an API response. A detailed examination of the topology of PDC JSON-LD metadata is presented next.

For the JSON-LD endpoints, any given API response will be composed of JSON elements that correspond to schema.org types and properties. In particular, each individual metadata record is structured and nested to represent a Dataset instance. Furthermore, if a response contains one or more metadata records, then each individual Dataset element will be contained within a wrapping itemListElement, where the set of itemListElements are contained within an itemList (i.e. an array of itemListElements). Therefore, it is necessary to understand the structure of the metadata within the API response, in order to be able to effectively extract content of interest. Figure 2 illustrates the nested topology of PDC JSON-LD metadata, with validated content shown on the left and raw JSON shown on the right.

One of the core motivations for providing these APIs, and other resources, relates to supporting knowledge mobilization objectives. In general, APIs necessarily meet the goal of broad data dissemination, which is central to knowledge mobilization. The specific APIs discussed here also serve to push appropriate data into the public domain, another important facet of knowledge mobilization. Beyond these features, we expect that these APIs will drive knowledge synthesis, reinterpretation, and subsequent redistribution, thereby extending the scope of dissemination. In particular, we are expecting that the API endpoints will be harvested and consumed by automated processes and scheduled tasks and that their contents will be used to populate other repositories and catalogues, but also as data sources for real-time data visualizations and machine learning applications.

Overall, the metadata APIs described in this section will benefit government and university researchers and scientists, as well as graduate and undergraduate students, who are involved in polar research. However, in addition to polar and geographic domains, we believe that these resources will also interest and facilitate access for a wider domain of data scientists and modelers, as well as the interested public. Indeed, one of the key goals of these ventures is to achieve knowledge mobilization by linking currently available polar research to broader arenas, where it can contribute to pressing large scale issues and help to inform action and policy.

Organizational Policy

Much of the previous discussion has focused on standards, architecture, and technical implementation. While the topic of policy, whether from an external user policy perspective or from an internal organizational perspective, seems to be unrelated to these previous topics, there is in fact considerable interplay among these facets. This is because policy choices inform downstream decisions on standards, architecture, and technical implementations, while policy implementations are necessarily bound the standards, architecture, and technical capabilities of the repository. That is, a repository cannot simply elect to support FAIR data principles if its current implementation lacks the capability to provide the corresponding functionality. Thus, understanding the resonance between these two key drivers of RDM is crucial and should be undertaken as early as possible in the lifecycle of any repository, preferably before its public launch.

PDC already has a Data Policy for metadata and data artifacts, as well as a general Terms of Use for the website, including the associated applications and databases. However, these constraints are predicated on a paradigm of manual usage, where a human user will need to click to assent to these usage agreements, in order to use the service. For the previously discussed APIs, the usage domain will necessarily need to support scenarios beyond manual activity, such as automated harvesting of metadata. Therefore, a different approach is required with respect to usage and agreements.

One approach would be to require API users to acquire an API key in order to be able to make calls to the metadata API. An API key acts as a unique identifier that is used to authenticate a user, and in addition ensure that the particular client has acquiesced to all applicable policies, licenses, and agreements, prior to the granting of the key. This would allow automated usage without manual acceptance of terms, because all terms have been previously accepted by the particular client. Another approach is to forgo repository-specific terms in favour of a public license, such as a Creative Commons license . Creative Commons licenses offer flexible features, including reuse and commercialization considerations, as well as supporting machine-readable metadata generation, intended to facilitate attribution of the licensed work. The Apache license, version 2.0 is another popular public license, but its usage is primarily intended for software licensing, while the Creative Commons licenses are more broadly applicable to creative works in general.

Understanding and determining the best suited policies and applicable licenses is mission critical for any data repository. These decisions become even more imperative when a repository requires user registration because changing, updating, or revoking existing agreements and licenses can become a nearly insurmountable issue. This is exacerbated by the definition of ownership, as data submitters typically maintain ownership of their data assets, although they typically have no dominion over the corresponding metadata that describes these artifacts. Thus, careful planning needs to go into repository policy design and the selection of corresponding licensing and agreements.

Modern trends in cloud infrastructure have given rise to another major policy choice for data repositories: Specifically, there now exists the opportunity to implement serverless architectures for selected services, or even the entire repository. The concept of a serverless architecture involves removing the concern of server management at the repository level and delegating this to a cloud services provider, like Amazon Web Services or Google Compute Engine. The motivation for this approach is that data repositories are then able to focus their technical efforts on domain-specific application and database development, without the considerable burden of maintaining physical infrastructure. However, this premise, while appearing to be purely of a technical nature, carries significant policy ramifications. One of the primary concerns relates to the geographic location of stored data. In particular, there can be caveats stemming from the transitive imposition of laws and regulations that are inherent to a given country, state, or other entity. This is often is major concern for submitters of data, and with good reason. Therefore, substantial diligence is required before adopting this type of design, in order to manage and mitigate organizational risk, as well as to understand potential changes in user patterns and demand. Thus, the decision to use a serverless architecture provides a compelling example of the powerful coupling between technical and policy perspectives.

Conclusions

Data repositories, such PDC, face a myriad of complexities and challenges, from both technical and non-technical perspectives. In addition to the ongoing evolution of new technologies and progressive changes in policy and goals, there is the complication of reconciling the dependencies that arise from the interplay between all of these factors. That is, the decision to implement FAIR data principles cannot be done in the absence of supporting technical infrastructure.

Similarly, electing to implement a metadata API while deprecated another older metadata harvesting service, such as an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) endpoint, cannot be done without assessing and understanding policy issues relating to potential impacts to users of the deprecated service.

Future plans for PDC include a full public release of our metadata API with support for legacy ISO 19115 XML, as well as schema.org compliant JSON-LD. In conjunction with the API release, PDC will also integrate its metadata to be available through the CCADI metadata API, to support its own forthcoming public release. PDC will also be updating, strengthening, and clarifying the content and language that comprise its Data Policy and Terms of Use. Specifically, the goal is to modernize the context for these documents so that they better reflect and support the interests and activities of the polar ROM community, and data management at large. PDC will also broaden its dissemination efforts beyond metadata to also include data by developing infrastructure to expose selected datasets through OPeNDAP services.

A key challenge for many data repositories, including PDC, will be to begin planning for the integration and adoption of Indigenous Knowledge, and more generally any cross-cultural knowledge, which is not yet directly supported. This undertaking serves as another paramount example of the reciprocal nature of organizational policy and technical implementation, with respect to their impacts upon one another. Methodical planning will be essential to effectively implement a design that is both powerful and sufficiently malleable, so as to meet current needs, while offering capabilities to meet future contingencies.