Toward Best Practices for Unstructured Descriptions of Research Data

Dan Phillips
Dalhousie University, Canada
DPhillips@dal.ca

Michael Smit
Dalhousie University, Canada
Mike.Smit@dal.ca

ABSTRACT

Achieving the potential of widespread sharing of open research data requires that sharing data is straightforward, supported, and well-understood; and that data is discoverable by researchers. Our literature review and environment scan suggest that while substantial effort is dedicated to structured descriptions of research data, unstructured fields are commonly available (title, description) yet poorly understood. There is no clear description of what information should be included, in what level of detail, and in what order. These human-readable fields, routinely used in indexing and search features and reliably federated, are essential to the research data user experience. We propose a set of high-level best practices for unstructured description of datasets, to serve as the essential starting point for more granular, discipline-specific guidance. We based these practices on extensive review of literature on research article abstracts; archival practice; experience in supporting research data management; and grey literature on data documentation. They were iteratively refined based on comments received in a webinar series with researchers, data curators, data repository managers, and librarians in Canada. We demonstrate the need for information research to more closely examine these unstructured fields and provide a foundation for a more detailed conversation.

KEYWORDS

Research data management; metadata; data documentation; data summarization; human data interaction

INTRODUCTION

To encourage data reuse and maximize benefits of conducted research, funding organizations around the world are implementing policies that require research data to be made publicly available following completion of a project (cOAlition S, n.d.; European Commission, n.d.; National Science Foundation, 2020).

There are three ways in which data are normally published and encouraged for reuse: data can be included within a scholarly publication (perhaps as an appendix), they can be published in a data repository as a discrete object, or they can be published in a data journal as a research output describing a dataset in complete detail (Schöpfel et al., 2020). Recognizing that one of the primary purposes of publishing research data is to facilitate reuse (Wilkinson et al., 2016), and that not every dataset can receive the time and attention necessary for data journal publication, the area of greatest growth is data repositories. Common data repositories for non-specialized environmental research include Data Dryad, Pangaea, Figshare, and Zenodo, as well as national, institutional, and discipline-specific repositories.

Research data is generally documented using controlled vocabularies and well-defined metadata fields, supporting precise search and discovery tasks (Shen, 2017; Chapman et al., 2020) and a machine-readable semantic web. They are outlined by standards (e.g., the Digital Documentation Initiative (DDI) Codebook) and are essential to sharing data across multiple repositories (Corti et al., 2014; FRDR, n.d.; Google, 2020). Yet most repositories retain natural-language fields for data description, including the title title and one or more blocks open-ended, unstructured text describing the data, often called the abstract, description, or summary. (We refer to this field as a dataset summary, without loss of generality.) The casual searcher is unlikely to choose search terms from a controlled vocabulary, and essential elements of the data collection are not captured in existing controlled vocabularies, so the title and dataset summary make routine search and discovery tasks possible (Chapman et al., 2020).

While the abstract section of a research article is well-understood and extensively studied, well-documented (from informal guides to ISO 214:1976), and a routine part of graduate student training, the dataset summary remains largely undocumented. Anecdotally, research data management (RDM) librarians suggest they build their expertise over years, and use that expertise to support individual researchers, similar to an archivist's expert approach to resource description. The guidance provided to scholars to assist in describing their own data in a way that supports search and discovery is limited, and review of how this field is used reveals an unsurprising lack of consistency and limited utility (see Literature Review and Environment Scan). As data repositories rise to the challenge of providing relevant search

84th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 3, 2021 | Salt Lake City, UT.

results from expanding collections, dataset summaries of consistent quality will be essential. Writing high-quality dataset summaries should be well-understood, well-documented, and routinely taught.

This paper first establishes the necessity of improving our approach to dataset summaries as a discipline, reviewing relevant literature and conducting and environment scan of how dataset summaries appear across some of the most popular data repositories. In the Literature Review and Environment Scan section, we examine how dataset summaries are solicited at the time of data submission, describe how dataset summaries appear when searching, and report the state of dataset summaries in key repositories. Our literature review makes it clear that this is an understudied area in the field of information management, and our environment scan describes the consequences of this neglect.

Second, we take initial steps toward these best practices. We examined existing best practices for writing summaries (paper abstracts, archival description, and much more) and synthesized candidate guidance for research dataset summaries. We drew on our combined expertise and years of providing support to researchers in managing research data to refine the guidance. We then conducted a series of webinars with expert audiences to disseminate our prototype guidance, which we iteratively improved and further refined in response to feedback from the data librarians, managers of data repositories, discipline-specific repositories, and researchers in attendance. We describe this in the Methods, Results, and Discussion sections.

LITERATURE REVIEW AND ENVIRONMENT SCAN

Abstracts for research articles underwent a shift, from professional abstractors to those provided by the author of the resource (Borko & Bernier, 1975; National Information Standards Organization, 2015; Tenopir & Jasco, 1993). Reviewing abstracts has always provided researchers with a briefing of current research, helped them determine whether it is worth resources to obtain and review a full article, and simplified complex topics (Tibbo, 1993). In the context of today's federated search tools, they provide an indexable field, aligned with natural-language queries, and a quick way for searchers to see their keywords used in context and adjust their query (Chapman et al., 2020). Most result pages for research article searches show the title of the paper and a portion of the abstract (or full paper), with the keywords shown in context.

Whereas decades of knowledge from professional abstraction have been transferred to article authors, data summaries have not had the same knowledge transfer. Research data librarians, some researchers, and some research staff have built expertise in writing data summaries, but there has been little effort to capture this expertise and transfer to others. Key questions remain open for discussion, suggesting no agreement on best practices. From the beginning of repositories, most data summaries have been written by their creator, rather than a summary specialist or abstractor, though perhaps with the support of an RDM librarian. Yet in our review of the literature and our environment scan of common data repositories, we found that very few cues are available to creators, with predictable results.

Data Description in the Literature

The study of summarizing spans research in education, psychology, linguistics, and others (e.g. Hidi and Anderson, 1986). Writing summaries for an audience (what is potentially valuable for others) is different from writing a summary for oneself (what is important to self), and that the former is substantially more difficult (Hidi and Anderson, 1986). While there are elements in common with other forms of summarization, the translation from structured data to a meaningful description is clearly different from the process of summarizing text (Koesten, 2018), for example including elements of information not actually present in the object being described; imagining the potential uses of data instead of reinforcing the key message and contribution; and more.

Koesten et al. (2020) suggested best practices for open data summarization from the standpoint of understanding and improving dataset search. They share our view that this area is understudied; the bulk of previous literature on data description is from authors associated with this group. In their 2020 study, they asked data-literate users on a crowd-sourcing site to describe sample data based on screenshots. Using constraints established in a pilot test, participants were shown sample summaries, limited to 50-100 words, and were given 3-12 minutes to write each summary. They coded those summaries to understand the high-level information and the attributes summarizers focused on. They found participants described the data structure, the scope of data through a focus on the variables present, the quality of the data (e.g. missing values), and some ideas for analysis or use of the data.

Based on their findings and a previous study (Kacprzak et al., 2019), they propose a 9-question dataset summary template for use during data ingestion. Our approach and the Koesten et al. (2020) approach are different lenses on the same problem, and we suggest that both approaches are required. Understanding how users find data is essential to writing good data summaries. Understanding the natural tendencies of humans when asked to summarize datasets

offers insight into how we should guide and shape this behavior when setting expectations. Yet time-limited, word-limited summaries from lay users are an incomplete source of information to inform data summary best practices. Our approach, to be guided first by authority and experience captured in written evidence, and then refined by a broad suite of stakeholders, runs the risk of under-valuing user experience, even as we focus on data descriptions written by data owners. In the Discussion section, we reflect on how our recommendations align with theirs, and in particular how their work answers an essential question that we raise.

Data Description Instructions

There is little instruction available to data creators on the contents of a data description. The most basic question, "what should it include", remains unanswered. Should the study that produced the dataset be described? The contents of the dataset, like key variables? The steps in processing the data? The file formats and standards used? All of the above? In what level of detail? We reviewed the interface and documentation available to data submitters / searchers at 5 major repositories (and Google Dataset Search). Table 1 shows the sparse information available.

The data submission interface is where many data submitters will start and finish in their search for instructions, but we must also consider data and metadata standards. Many are intended for expert users, and are not detailed. The Dublin Core Description element referenced is documented sparsely: "an account of the resource". A comment elaborates that "description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource". The most detailed formal documentation we were able to locate was in the DDI Codebook, which we reproduce here in its entirety:

> An unformatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs [principal investigators] attempted to answer when they conducted the study. A listing of major variables in the study is important here. In cases where a codebook contains more than one abstract (for example, one might be supplied by the data producer and another prepared by the data archive where the data are deposited), the "source" and "date" attributes may be used to distinguish the abstract versions. Maps to Dublin Core Description element. Inclusion of this element in the codebook is recommended. The "date" attribute should follow ISO convention of YYYY-MM-DD. (Data Documentation Initiative, n.d.)

This advice, intended for expert users and not readily available to data depositors, is still limited in detail. There is no guidance for language, length, or intended audience, and the terms used ("major", "special characteristics") are ambiguous. 13% of it (by word count) references other standards. This is sufficient documentation to start the task, but not sufficient to finish.

Almost two-fifths of the DDI entry refers to possibility of multiple summaries for a single data deposit, which is also considered a best practice in other contexts (Bascik et al., 2020) but is uncommon in actual practice. Having multiple summaries could speak to different levels of technical understanding, provide overviews for people who speak other languages, or give updated descriptions for datasets which change over time. Other guides have suggested using separate fields for the abstract and the methodology (DataCite Metadata Working Group, 2016), which Pangaea has adopted (among very few others).

Platform	Default summary-writing instructions
Dataverse (Scholars Portal, Canada)	Description: A summary describing the purpose, nature, and scope of the dataset.
Data Dryad	Abstract: Short description of dataset.
Figshare	Description: Describe your data as well as you can. Formatting is preserved when pasting from other sources and counts towards character limits.
Pangaea	Description: ABSTRACT and/or further details describing the data.
Zenodo	Description
Google Datasets	Description: A short summary describing a dataset.

Table 1. The instructions provided on what to enter in the data description field provided by various data repositories and indices as of August 2020.

Data Description Exemplars

One way research article authors learn to write abstracts is by imitation: research training and research work necessarily includes exposure to hundreds or thousands of articles, with their abstracts. A search for journal articles shows portions of abstracts. Aspiring authors will find samples of great abstracts in every imaginable field a mere Google search away. In contrast, the lack of agreement on best practices for dataset summaries limits the number of positive examples. RDM experts tell us they know a good description when they see one, but this knowledge is not transferred to data creators and owners other than through iterative feedback.

We observed common data repositories for the presence or absence of key best practices for showing resource summaries in search results, and discovered that most have not adopted the practice of showing abstracts in their search results, and instead rely on the dataset title, as shown in Figure 1 and Table 2.

Note that both Dataverse and CKAN are open-source platforms that are highly customizable, and results may vary depending on the implementation chosen. (Open Data Canada is exclusive to government data, and the data submission page is not available to us.) These two platforms made the best use of dataset summaries.

Data Description in Practice

A systematic and comprehensive examination of dataset summaries is not possible without widely-accepted standards on how to write a good dataset summary. (In contrast, examination of the quality of journal article abstracts is a routine and ongoing practice, e.g. Wang et al., 2017; Nagendrababu et al., 2019). We examined dataset summaries across multiple repositories to build our own assessment of the state-of-the-art, and found that quality is (predictably) highly varied. We describe here some quantitative indicators of the state of dataset summaries.

Figshare is widely used for disseminating research artifacts, including datasets. Their dataset summary minimum length is 4 characters, and there exist datasets with summaries of this approximate length. We found dataset summaries that consist entirely of a list of the file formats available. Summaries that are shorter than the title are surprisingly common. At the other end of the spectrum, we found datasets carefully and meticulously described.

Deposited data are often supplement to journal articles, sometimes through automated processes that connect journal hosting systems with data repositories. In some research data repositories, it is normal to copy article abstracts verbatim to their supplementing dataset. This is true of over 36% of examined dataset summaries in Figshare (n=278 in a sample taken for this environment scan) and 95% of data summaries in Data Dryad (n=338). In many situations, common procedure leads to best practice – but the purpose of an article abstract is to describe an article and will include results based on the analysis of the data, which will not be apparent from the data alone. Even based on the brief documentation in DDI, this approach is not appropriate for a dataset summary, yet it is common.

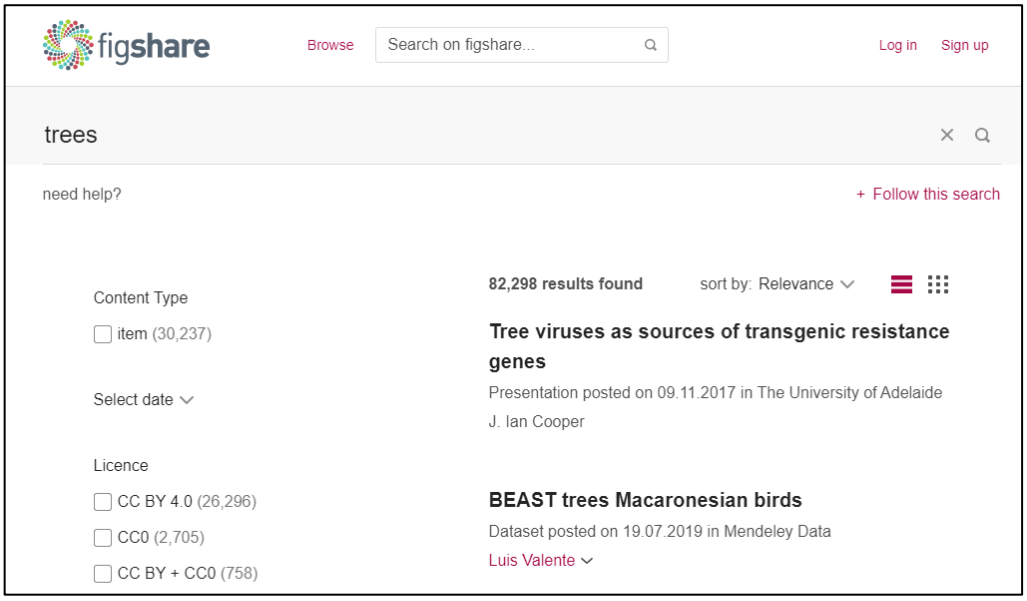

Figure 1. The search results page for figshare after a search for "trees", limited to datasets. The display prioritizes title, date, and author, rather than showing a dataset summary with keywords in context.

Platform	Observation
Data Dryad	No dataset summary shown on search results page.
Figshare	No dataset summary shown on search results page.
Pangaea	No dataset summary shown on search results page.
Zenodo	Dataset summary shown when available (dataset summaries are optional)
Dataverse (Scholars Portal, Canada)	An excerpt from the data summary is displayed; search terms are emphasized.
CKAN (Open Data Canada)	A complete summary is displayed; search terms are emphasized.

Table 2. The presence or absence of dataset summaries on the search results page of widely-used data repositories, as of August 2020.

METHODS

We developed the proposed best practices for data summarization following a two-phase approach. Phase 1 synthesized existing knowledge on summarization into candidate principles, which underwent iterative review and discussion within the research team based on our experience supporting RDM. Phase 2 was intended as a knowledge transfer activity, sharing the results of our environment scan and literature review with key stakeholders, with beta concepts for effective data summaries to spark discussion. For this candidate version, we have incorporated and reflected on feedback from data and metadata experts, data curators, librarians, researchers, and research support staff. An overview of our approach is shown in Figure 2, and each phase is described in more detail below.

Phase 1

Key pieces of advice were extracted from a variety of sources related to summarization, including literature related to abstracting techniques for research articles, search strategies, the use of language, metadata, and data repositories. We also drew extensively on grey literature, including best practices from other disciplines, such as library catalogues for specific disciplines (especially health sciences); Rules for Archival Description (RAD); widely accepted principles such as the FAIR (Findable, Accessible, Interoperable, and Reusable) Principles for Scientific Data Management; metadata standards, especially those intended for data; and journal, repository, and federated search technical documentation for metadata. Much of this original source material was not directly applicable but could be translated to the context of describing research data. These candidate pieces of advice were reviewed, and those not applicable were discarded (filtering). Those that remained were iteratively sorted into loosely defined themes such as provenance and structure (aggregation). From this, a draft list of concepts for effective data summarization emerged and was used to intuitively rewrite existing data summaries to attempt to put the concept into practice. (One such revised data summary is in Figure 3.) Discussion and reflection within the research team served as an initial quality review.

Phase 2

In our effort to spark a wider conversation about this subject, we took a preliminary version of our guidance to a series of online, synchronous, video conference webinars. We held four private webinars with key working groups in the research data management space, and one public webinar advertised to the research data management community in Canada. Participants in the webinar were advised that we were there to share our work, but were also interested in their feedback, which we would use to improve our work.

Each of the working group consultations began with a ten-minute presentation outlining the current state of data summaries in public data repositories. This was followed by a twenty-minute moderated discussion focused on the concepts for best practices. Each working group had between seven and seventeen members. Discussion notes were taken at each meeting by the researchers. The research team met after each to share their notes and discuss the comments received to iteratively refine the guidelines. The final improved guide is described in the following section.

The first two consultation groups were from the Canadian Integrated Ocean Observing System (CIOOS; Stewart et al., 2019). The first of these was with a regional cross-section of CIOOS stakeholders, including oceans researchers, technical experts, user engagement specialists, and research administrators. The second was with a national committee of technical experts tasked with building the CIOOS repository, including the submission form, the onboarding process, and the standards, metadata, and controlled vocabularies in use. (Disclosure: several members of the first group report to one of the co-authors; the second group is co-chaired by one of the co-authors.)

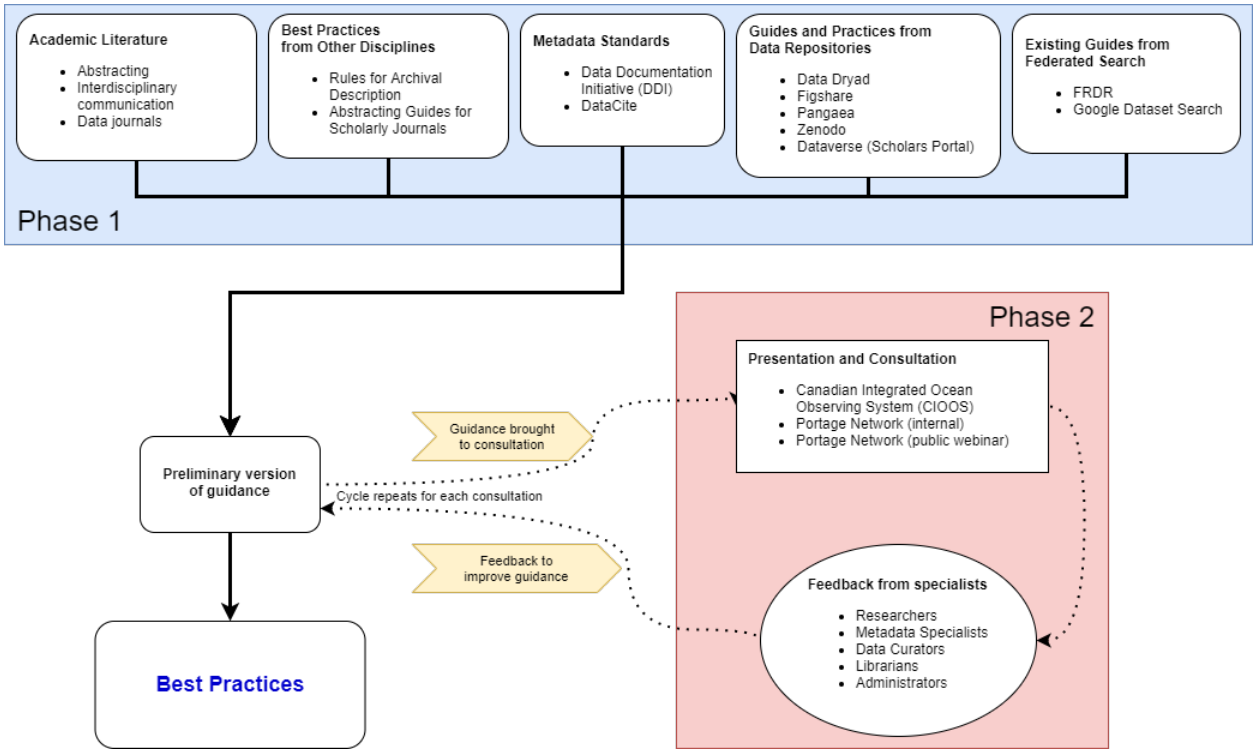

Figure 2. Visual summary of the methods used to create our best practices for dataset summaries.

The next two consultations were standing expert groups at the Portage Network for Research Data Management, the data stewardship body of the Canadian Association of Research Libraries (CARL), and part of the New Data Research Infrastructure Organization (NDRIO) in Canada. One group focuses on how users search for and discover data; the other on how to submit, document, and curate data. Both groups are national in scope, and include metadata specialists, RDM librarians, archivists, systems experts, and others, who are invited from across Canada based on their expertise.

The final public expert consultation was a similar format, going into greater detail on the existing nature of public repositories and issues with their data summaries. The webinar ran for over 50 minutes equally divided between presentation and discussion. It was attended by 52 live participants who were not screened for expertise (a recording of the session is publicly available on Youtube: https://www.youtube.com/watch?v=kUIoX3OB130).

RESULTS

Our method of identifying key advice and filtering resulted in some very specific advice, but the second step (aggregation) ensured that advice remained high-level. These best practices are not yet suitable for teaching or training researchers to write dataset summaries, or for assessing quality. Rather, they serve as a starting point for discipline- or context-specific guidance. More specific guidance (word counts, structure, vocabulary, headings, and other such matters) will, like research article abstracts, be guided by the relevant scholarly community, publishers, and the experience of both creators and searchers. Our conclusion is that data summaries vary so widely in style and function that we do not yet have a clear idea of what specific elements are effective, but that ultimately variation among disciplines is likely necessary. These best practices are a starting point; if followed, writers of data summaries will develop norms from which effective techniques can be identified and best practices can continue to evolve.

We describe here five high-level principles intended to guide the establishment of discipline-specific guidance for dataset summaries. For each, we describe the provenance, variations and changes considered, and an explanation of the importance. Without loss of generality, we focus our examples on environmental data.

Best Practice 1: Describe Datasets with Attention to the Searcher

Convenience of the user is the first and most important of the International Cataloguing Principles (Galeffi et al., 2016, p. 5) and has been adopted in some form by several related standards for description (Bureau of Canadian Archivists,

1990; Wilkinson et al., 2016). When presented with the description, a searcher should be able to quickly identify whether a resource is *useful* to them and whether the resource *can be used* by the searcher.

To determine whether a resource is *useful*, a searcher should be able to identify whether data are relevant to their needs. The quality of relevance has been extensively studied in information retrieval research, which we do not reproduce here. We simplify it to this: the searcher of environmental data will have a scope for their search that overlaps with one or more dimensions of the dataset. Dimensions common to environmental data include geospatial (the location where data were collected and/or the location that is immediately impacted by collection), temporal (the date range that the data covers), and subject. While structured metadata can retrieve data by dimension, the decision on relevance is ultimately made by humans, and the dataset summary is essential to this task.

To determine whether a resource *can be used*, a searcher should be able to identify the extent to which data have been processed and the format the data are in. A dataset summary should not be so specific as to give file formats that may change with versioning (*Data on the Web Best Practices*, n.d.), but it should be possible to provide an understanding of data structure such that users will know whether they have the technical expertise to use them (Wilson et al., 2017).

There was some discussion within Phase 2 on whether restrictions on data use (i.e., licensing and rights considerations) are appropriate within the data summary. If so, this is arguably a consideration for whether data *can be used*. We decided to not recommend this information be included, as permissions for use may change with time. While licensing information should always accompany a dataset, it does not need to be contained within the summary.

Best Practice 2: Begin with Language Appropriate to all Potential Audiences

For both expert and less-experienced users, terminology is not always understood outside of a specific discipline or the context of an in-depth analysis (Montesi & Urdiciain, 2005). It is generally understood that a summary should be written such that a layperson can understand it (Borko & Bernier, 1975).

Writing to the lay reader has been a specialized skill that professional abstractors have developed. Although subject matter specialists writing their own summaries may not have developed this skill (Borko & Bernier, 1975), this still provides an objective to strive towards. The summary should begin such that readers understand the general topic of a dataset from reading the first few sentences, and depending on their literacy of the topic, continue reading to understand it in greater depth. In situations where data cannot be sufficiently described without the use of jargon, it might be appropriate to incorporate impact statements for laypersons (Sick, 2009) as an introduction to the summary.

We stop short of requiring that the entire summary be appropriate to all audiences. One portion of the audience, particularly in the research data context, is expert users. Their determination of relevance will rely on the details and vocabulary of domain experts. For example, a climate modeller accessing ocean observation data will require detailed and specific information about how the water salinity was measured to ensure consistency, comparability, and quality. While much of this detail is appropriate for accompanying documentation, and the use of a controlled vocabulary for variable names, this expert user will be well-served by the use of the terms “refractometer” or “absolute salinity” in the summary, as they imply a wealth of information. However, this detail belongs later in the summary, and it requires judgment to use technical jargon in appropriate measure.

Best Practice 3: Describe the Dataset as an Independent Research Output

In response to the concern that scholarly journal abstracts are copied directly into the summaries of the datasets that supplement them, describing the dataset as an independent research output reinforces the idea that a dataset can be (and should be) considered a standalone object. Other description standards speak to this through a statement on accuracy, describing discrete items or collections without describing those around them (Bureau of Canadian Archivists, 1990, p. 23; Galeffi et al., 2016, p. 5). The principle of sharing research data suggests that the data has value beyond the work already done: while some of this work may be replication, other types of analysis are possible. The data should be described appropriately.

Data which supplement a resource (such as an article) should still be linked to that resource. This could be done either within the data summary or through the use of another user-facing metadata field. It is important that the description of the source study not overtake the description of the dataset as a discrete object. The practice of copying abstracts may benefit user groups who are searching for similar studies, but the link between research artifacts should not rely on the similarity of unstructured text fields. Relying on a research article abstract will obscure datasets from other searchers who could benefit from the reusability of data in less direct ways.

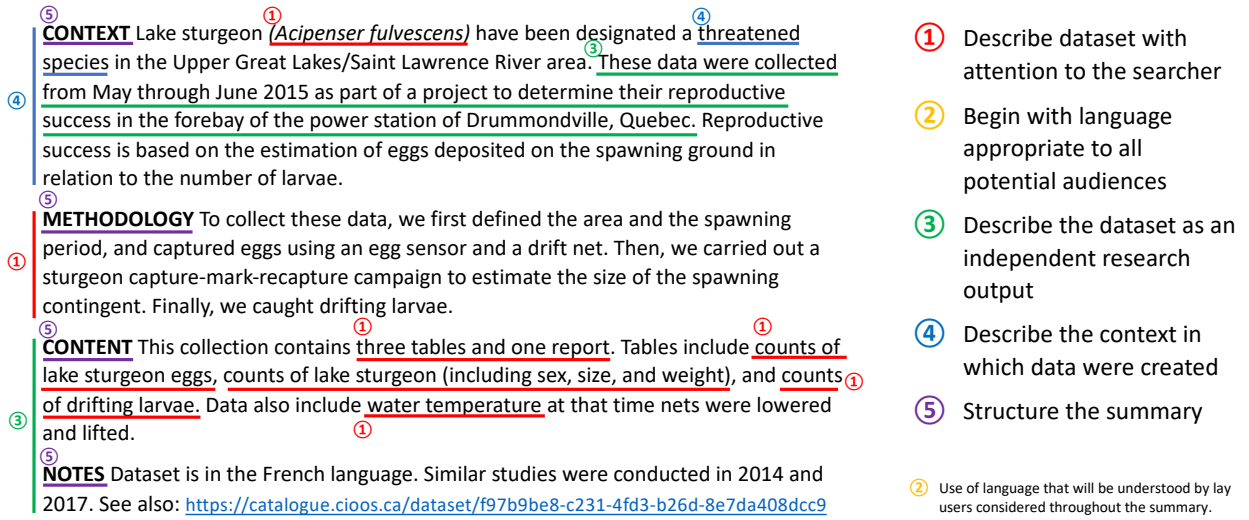

Figure 3: A sample dataset summary written following the best practices proposed here, annotated with numbered and color-coded notes to illustrated how the best practices guided language. Source: (SLGO, 2015)

Best Practice 4: Describe the Context in which Data were Created

While datasets can be considered an independent research outputs [Best Practice 3], environmental data are not created without an intended purpose. Understanding the original context for their creation is necessary for evaluating provenance, the completeness of data, and the degree to which they have been processed. Looking to the exemplar data summary (Figure 3), a dataset has been published out of a project to measure the reproductive success of lake sturgeon *(Acipenser fulvescens)*. Without understanding that lake sturgeon is designated as a threatened species in the area where these data were collected, the conditions described by these data may be misinterpreted as typical.

Put another way, this best practice asks researchers to provide the "why" in addition to the "what". The data summary should explain what is interesting, unique, important, notable, or otherwise relevant to understanding the dataset.

Best Practice 5: Structure the Summary

Structured abstracts with consistent headings are considered more readable and help to provide consistency across publications. These are common in fast-moving disciplines such as medicine, and there are efforts to bring this practice into other areas of research (Hartley, 2014; Hartley et al., 1996; Montesi & Urdiciain, 2005).

In the medical sciences, a structured article abstract consists of clearly labelled sections such as Background, Aims, Participants, Methods, and Results. Using a similar structure for environmental data can help develop search tools to distinguish background from methodology (DataCite Metadata Working Group, 2016), which in turn helps address the context of creation [Best Practice 4] and help researchers find studies that use similar techniques and tools. During Phase 2, one consultation group highlighted using structure as means of addressing reuse potential or concerns with data which are necessary in their interpretation (i.e., flagging missing data). This can be compared to the way a scholarly article abstract should be clear about the limits of the study. One possible structure is shown in Figure 3.

DISCUSSION

Our work was based on our understanding of how users search. There is some indication that dataset search has unique properties (Kacprzak et al., 2019), but our fundamental assumption is that search behavior is learned, and that users seek to apply skills learned in one type of search to other contexts. This informed our strategy to begin with literature and practices from other forms of catalogues, and to compare data repository search interfaces with academic journal databases. Several data discovery experts in our consultations believe that this is a flawed assumption: that researchers often have a very clear understanding of what usable data should exist and refine their search criteria using clearly defined filters rather than broad search terms. One participant in Phase 2 commented that the summary is the very last thing that a researcher in their field would read. While this type of search filtering remains essential, we continue to assert the importance of the dataset summary, based on our confidence in information retrieval literature and mental model user experience research (supported by Koesten et al., 2020). In general, our reflection on input was guided by our recognition that experts do not always see the perspective of end users.

One challenge to our approach is that data summaries are often not written with the searcher in mind. People who are experts in search and retrieval may have developed other techniques to overcome the limitations of poorly written data summaries. It would be interesting to see a comparison of the processes used by experienced data searchers with researchers who are searching for data for the first time. This was discussed within the public consultation, noting that data repositories used by academic institutions have a very different audience, and they often have research experience already. This is different than government open data portals in which the inexperienced public are encouraged to browse for interesting data. The techniques used by each audience may have led to the different ways that repositories display data summaries in search result pages (see Environment Scan). Likewise, different search techniques may be used between people interested in environmental resources (often natural scientists) and other groups seeking characteristics of an environment to supplement social science observations. The importance of discipline expertise in reviewing, implementing, and developing processes based on our best practices is essential.

There was a recurring question on the notion that it is always possible to write summaries that any reader can understand, or that this should even be a primary objective. Some searchers *are* experts, and in many contexts, these are the users who can best reuse data. The prevailing belief is that experts have other means of finding data in the current search environment without using data summaries at all. Including our best practice – that summaries *begin* with language appropriate to all audiences – suggests that expert language can be used as the summary progresses while still providing a foundation for those who do not yet know whether a particular deposit is useful.

We also considered the specific questions suggested by Koesten et al. (2020), listed in Table 3, in light of our high-level best practices. We found common themes. In general, the user-focused methods employed by Koesten et al. (2020) align with our Best Practice #1 (BP1): they have undertaken substantial user studies that provide concrete information to inform the implementation of BP1. Questions 2, 3, 4, 5, 6, and 7 serve this role: providing sufficient information to enable the searcher to assess relevance. We would suggest that 5-7 should not be optional: while recognizing that setting achievable goals is valuable, we should aim for the data summaries we want to see. Question 8 aligns with BP 5, where we suggest that understanding limitations of the data is one benefit of structured abstracts. This suggests that we should be including limitations in our assessment of useful / usable headings. Structured data summaries are also aligned with their approach of asking specific questions to guide the creation of dataset summaries.

Template question	Explanation
*1. How would you describe the dataset in one sentence?	What is the dataset about?
*2. What does the dataset look like?	File format, data type, information about the structure of the dataset
*3.What are the headers?	Can you group them in a sensible way? Is there a key column?
*4. What are the value types and value ranges for the most important headers?	Words/numbers/dates and their possible ranges
5. Where is the data from?	When was the data collected/published/updated? Where was the data published and by whom? (required if not mentioned in metadata)
6. In what way does the dataset mention time?	What timeframes are covered by the data, what do they refer to and what is the level of detail they are reported in? (e.g. years/day/time/ hours etc.)
7. In what way does the dataset mention location?	What geographical areas does the data refer to? To what level of detail is the area or location reported? (E.g. latitude/longitude, streetname, city, county, country etc.)
8. Is there anything unclear about the data, or do you have reason to doubt the quality?	How complete is the data (are there missing values)? Are all column names self explanatory? What do missing values mean?
9. Is there anything that you would like to point out or analyse in more detail?	Particular trends or patterns in the data?

Table 3. The 9 questions (with documentation) proposed by Koesten et al. (2020) as a data summary template. The four required questions are noted with an asterisk.

There are also points of disagreement, which are not meant to detract from the important work of Koesten et al. (2020). Koesten et al. start from assumption that the summary author is not a domain expert, while our best practices are meant to guide data owners and other experts in writing effective summaries. Question 1 is suggesting a topic sentence, which is good advice, but we are concerned that without the expectation of BP2, the requirement for brevity will result in the use of technical language or an otherwise dense sentence. Question 9 suggests the summary include information on analysis of the data, which seems to contradict the spirit of BP3, to not pre-suppose the possible analysis of the data, and to not limit the summary to the use identified by the creator. It might also be seen as aligned with BP4, in relating the context. Finally, their advice is actionable, while ours is formative.

Who is responsible for authoring the dataset description remains an important question, with no clear answer. We think it is unavoidable that research data creators will bear this responsibility, but recognize that this may require thinking about more than improving data summaries: the entire data submission system may need to be rethought (Smit et al., 2020).

CONCLUSION

Research data management is an important sub-field in information research, building on concepts borrowed from archives, reference services, and data analytics. We must also borrow from information retrieval and user experience research to ensure that we realize the benefits of open research data.

Data repository platforms have a role in encouraging the ongoing development and enforcement of best practices. This is well understood by the working groups who acted as consultants throughout Phase 2: they exist to work with data repositories and data catalogues to promote and advance data stewardship. While many standards organizations provide frameworks and advice on how metadata should be developed, it is repository platforms that can best educate researchers on the expectations of their descriptions. This can be done in any of several ways, whether a callout box appears when a researcher begins to make a deposit, exemplars are provided for reference, a trained manual reviewer speaks with researchers, or the search experience passively highlights examples of well-crafted metadata. Of course, this advice must first exist.

The high-level best practices proposed in this paper are a step toward better data summaries. To be impactful, they must be matched with continued user studies, ongoing reflection and conversation, the development of discipline-specific guidance, and implementation in data repositories. In short, they must become more actionable, which requires continued conversation among interested scholars. Our ongoing and future work includes users studies; examining the impact of automatically translated data summaries; improving the semi-automatic generation of data summaries from structured metadata; and (carefully) crowd-sourcing the structured and unstructured description of datasets from the scientific community to address concerns about researcher burden and expertise.

ACKNOWLEDGMENTS

Research funding was provided by the Ocean Frontier Institute, through an award from the Canada First Research Excellence Fund; by a MITACS Research Training Award; and by the MEOPAR NCE Observation Core.

Some of the earliest conversations on this topic were with Adrienne Colborne (School of Information Management, Dalhousie University), Sarah Stevenson (Dalhousie Libraries), and Erin MacPherson (Dalhousie Libraries) in 2018. Thank you!

REFERENCES

Bascik, T., Boisvert, P., Cooper, A., Gagnon, M., Goodwin, M., Huck, J., Leahey, A., Steeleworthy, M., & Taylor, S. (2020). *Dataverse North Metadata Best Practices Guide: Version 2.0*. https://doi.org/10.14288/1.0388724

Borko, H., & Bernier, C. L. (1975). *Indexing concepts and methods*. New York : Academic Press.

Bureau of Canadian Archivists (Ed.). (1990). *Rules for Archival Description* (Revised Edition 2008). Bureau of Canadian Archivists.

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L. D., Kacprzak, E., & Groth, P. (2020). Dataset search: a survey. *The VLDB Journal*, *29*(1), 251-272.

cOAlition S. (n.d.). *Principles and Implementation | Plan S*. Retrieved December 14, 2020, from https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: a guide to good practice*. Sage.

Data Documentation Initiative. (n.d.). *XML Schema*. XML Schema Tag Library -- Version 2.1. Retrieved December 14, 2020, from https://ddialliance.org/Specification/DDI-Codebook/2.1/DTD/Documentation/abstractType.html

Data on the Web Best Practices. (n.d.). Retrieved December 14, 2020, from https://www.w3.org/TR/dwbp/

DataCite Metadata Working Group. (2016). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.0* [Application/pdf]. 45 pages. https://doi.org/10.5438/0012

DCMI Usage Board. (2020). Dublin Core Metadata Initiative (DCMI) Metadata Terms. Available: http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/

European Commission. (n.d.). *Data management—H2020 Online Manual*. Retrieved December 14, 2020, from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

FRDR. (n.d.). *FRDR Documentation*. Retrieved December 14, 2020, from https://www.frdr-dfdr.ca/docs/en/about/

Galeffi, A., Bertolini, M. V., Bothmann, R. L., Rodríguez, E. E., & McGarry, D. (2016). *Statement of International Cataloguing Principles*.

Google. (2020). *Dataset | Google Search Central*. Google Developers. https://developers.google.com/search/docs/data-types/dataset

Government of Canada. (2018). *DRAFT Tri-Agency Research Data Management Policy For Consultation—Science.gc.ca*. Innovation, Science and Economic Development Canada. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97610.html

Hartley, J. (2014). Current findings from research on structured abstracts: An update. *Journal of the Medical Library Association : JMLA*, *102*(3), 146–148. https://doi.org/10.3163/1536-5050.102.3.002

Hartley, J., Sydes, M., & Blurton, A. (1996). Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science*, *22*(5), 349–356. https://doi.org/10.1177/016555159602200503

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, *56*(4), 473-493. https://doi.org/10.3102/00346543056004473

Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., & Simperl, E. (2019). Characterising dataset search-An analysis of search logs and data requests. *Journal of Web Semantics*, *55*, 37-55. https://doi.org/10.1016/j.websem.2018.11.003

Koesten, L. (2018). A user centred perspective on structured data discovery. In *Companion Proceedings of the The Web Conference 2018* (pp. 849-853). https://doi.org/10.1145/3184558.3186574

Koesten, L., Simperl, E., Blount, T., Kacprzak, E., & Tennison, J. (2020). Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies*, *135*, 102367. https://doi.org/10.1016/j.ijhcs.2019.10.004

Montesi, M., & Urdiciain, B. G. (2005). Abstracts: Problems classified from the user perspective. *Journal of Information Science*, *31*(6), 515–526. https://doi.org/10.1177/0165551505057014

National Information Standards Organization. (2015). *ANSI/NISO Z39.14-1997 (R2015) Guidelines for Abstracts*. https://www.niso.org/publications/ansiniso-z3914-1997-r2015-guidelines-abstracts

National Science Foundation. (2020). *Dissemination and Sharing of Research Results*. National Science Foundation. https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nagendrababu, V., Duncan, H. F., Tsesis, I., Sathorn, C., Pulikkotil, S. J., Dharmarajan, L., & Dummer, P. M. H. (2019). Preferred reporting items for systematic reviews and meta-analyses for abstracts: best practice for reporting abstracts of systematic reviews in Endodontology. *International Endodontic Journal*, *52*(8), 1096-1107. https://doi.org/10.1111/iej.13118

Nature Research Journal Submission Guidelines. (n.d.). Retrieved January 15, 2018, from https://www.nature.com/sdata/publish/submission-guidelines

Schöpfel, J., Farace, D., Prost, H., and Zane, A. (2019). Data papers as a new form of knowledge organization in the field of research data. 1*2ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?,* ISKO France, Montpellier, France. ffhalshs-02284548f

Shen, Y. (2017). Data Discovery, Reuse, and Integration: The Perspectives of Natural Resources and Environmental Scientists. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–2. https://doi.org/10.1109/JCDL.2017.7991596

Sick, L. (2009). *Record structure for APA databases*. http://www.apa.org/databases/training/record-structure.pdf

SLGO. (2015). Assessment of the reproductive success of lake sturgeon at the Drummondville spawning ground in spring 2015 [data file]. https://catalogue.ogsl.ca/en/dataset/6c10db51-19a2-4a26-81dd-4464f295fb32

Smit, M., Ianta, A., and MacNeill, A. (2020). What if we reconsidered how we ask scientists to share their data: When FAIR meets crowd-sourcing and nudge theory. In AGU Ocean Sciences Meeting. https://doi.org/10.1002/essoar.10502628.1.

Stewart, A., DeYoung, B., Smit, M., Donaldson, K., Reedman, A., Bastien, A., ... & Whoriskey, F. (2019). The development of a Canadian integrated ocean observing system (CIOOS). *Frontiers in Marine Science*, *6*, 431. https://doi.org/10.3389/fmars.2019.00431

Tenopir, C., & Jasco, P. (1993). Quality of Abstracts. *Online Vol. 17*. https://trace.tennessee.edu/utk_infosciepubs/128

Tibbo, H. R. (1993). *Abstracting, information retrieval, and the humanities*. Chicago : American Library Association.

Wang, M., Jin, Y., Hu, Z. J., Thabane, A., Dennis, B., Gajic-Veljanoski, O., ... & Thabane, L. (2017). The reporting quality of abstracts of stepped wedge randomized trials is suboptimal: A systematic survey of the literature. *Contemporary clinical trials communications*, *8*, 1-10. https://doi.org/10.1016/j.conctc.2017.08.009

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, L., Colborne, A., & Smit, M. (2017). Preparing data managers to support open ocean science: Required competencies, assessed gaps, and the role of experiential learning. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 3984-3993). IEEE. https://doi.org/10.1109/BigData.2017.8258412