



How To Write A (Good) Data Description: Developing Best Practice

Michael Smit
Dan Phillips

 0000-0002-2028-4317
 0000-0002-0386-0132

Background

To encourage data reuse, research funding organizations around the world are implementing policies that require data management plans. As a result, the way we store data is getting more attention than ever before. Increasingly, researchers are looking at deposited data to use as examples when they create their own metadata. This includes the longform description.

Are we any good at writing longform summaries?

Repositories with the lowest barrier to entry are likely to get the most attention. These include Dryad, Zenodo, Pangaea, and institutional repositories such as Dataverse. In a scan of data summaries in these platforms, several reoccurring customs can be identified that may be perceived as problems. Consider:

- Metadata standards such as DDI and DataCite are often poorly followed in longform (uncontrolled) text fields. (They are understandably more difficult to enforce than other fields.)
- Descriptions are inconsistent in terms of length, depth, and level of technical writing. This may stem from researchers lacking experience reading other data summaries, as repository search interfaces make it difficult to see several summaries at once.
- Descriptions are often copied verbatim from article abstracts to which they are supplement. This does not explicitly describe the data.

Objective

Other disciplines such as archives, journals, and library catalogues have developed best practices for writing descriptions for effective search and retrieval. To address concerns in description quality, can we apply the best practices of other disciplines to the description of datasets?

Best Practices

These best practices have been developed through a review of literature and consultation with experts. These are deliberately high level and abstract, as all data summaries should find them to be relevant. Best practices for data description are proposed as follows:

1. Describe the Dataset with Attention to the Searcher

Convenience of the user is the first and most important of the International Cataloguing Principles and has been adopted in some form by several related standards for description. When presented with the description, a searcher should be able to quickly identify whether a resource is *useful* to them (content) and whether the resource *can be used* by the searcher (technical ability).

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. It is generally understood that a summary should be written such that a layperson can understand it. Where this is not adequate, start with a broad description and hone in on the details. An interested user will take the time to read it if it seems relevant.

3. Describe the dataset as an independent research output

In response to a 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. Data which supplement a resource (such as an article) should still be linked to it in some way.

4. Describe the context in which data were created

While datasets can be considered an independent research outputs, 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.

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.

Methodology

Phase 1

Key phrases related to summarization techniques were extracted from a variety of sources including academic literature, guidelines for abstracting, and principles for description in various fields. These were reframed to fit the context of data. For example,

“Creators of archival material must be described”
(Bureau of Canadian Archivists, 1990)

became

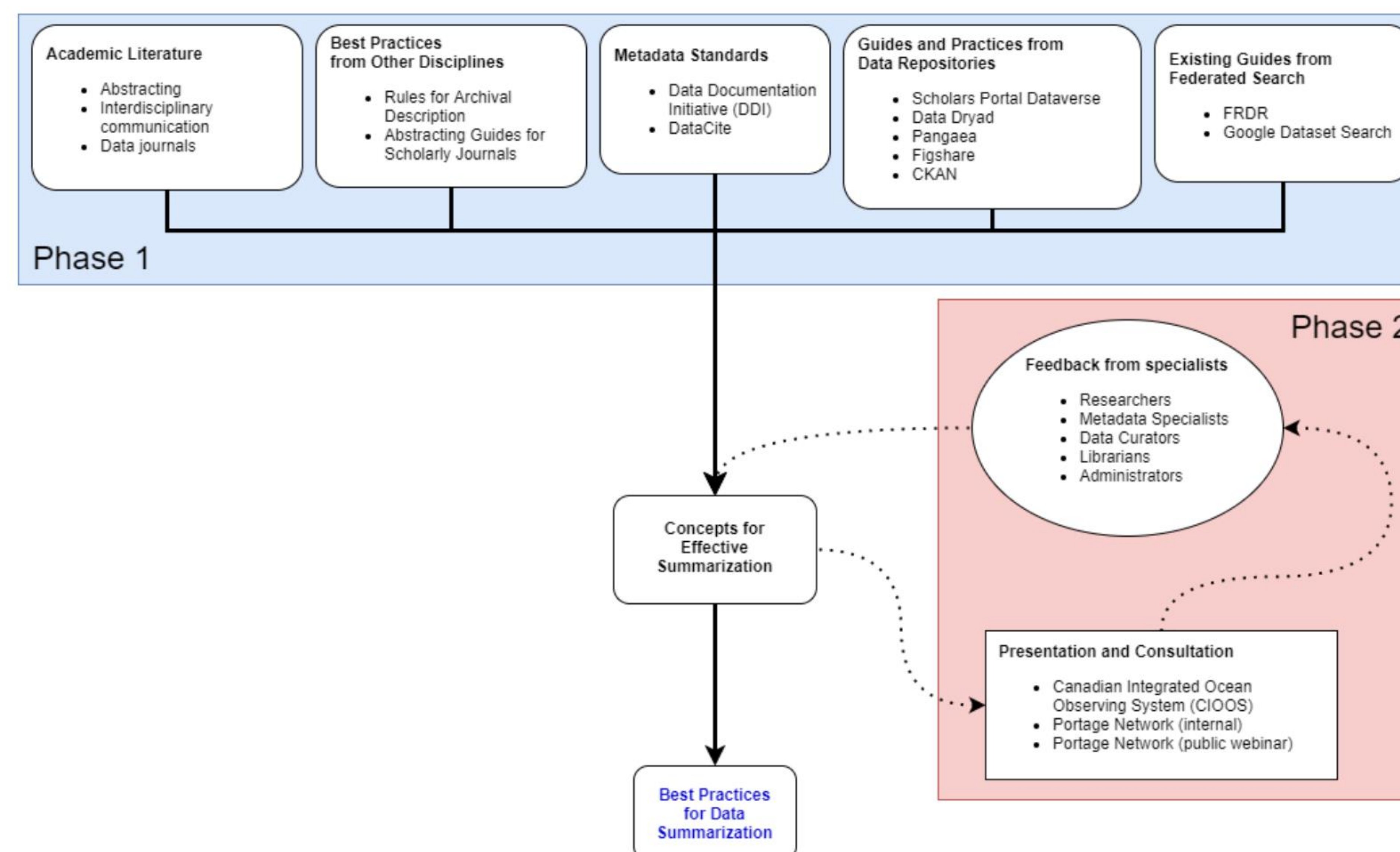
“The original purpose for data collection must be made clear.”

Similar statements were grouped into discrete categories. From this, concepts for effective summarization emerged.

Phase 2

These concepts were tested with expert consultation including two working groups within the Canadian Integrated Ocean Observatory System, two working groups within the Portage Network, and a public consultation hosted by the Portage Network. Perspectives included people who specialize in backend metadata development, people who specialize in frontend reference services, and people who work in systems integration.

Methodology Visualized



References

- Bascik, T., Boisvert, P., Cooper, A., Gagnon, M., Goodwin, M., Huck, J., ... Taylor, S. (2020, February 29). Dataverse North Metadata Best Practices Guide : Version 2.0 [R]. doi: <http://dx.doi.org/10.14288/1.0388724>
- Cory Chapman. (n.d.). *Usability Recommendations* [Wiki]. Datadryad.Org. Retrieved June 24, 2020, from http://wiki.datadryad.org/Usability_Recommendations
- Dryad. (2019). *Dryad Submission Process*. https://datadryad.org/stash/submission_process
- The Dataverse Team. (2020, April). *Dataset + File Management*. Dataverse. <http://guides.dataverse.org/en/latest/user/dataset-management.html>
- Google. (n.d.). *Dataset | Search for Developers*. Google Developers. Retrieved June 24, 2020, from <https://developers.google.com/search/docs/data-types/dataset>
- Hartley, J., & Sydes, M. (1996). Which layout do you prefer? An analysis of readers' preferences for different typographic layouts of structured abstracts. *Excerpt from Their Structured Abstracts in the Social Sciences. British Lib. 1995, 22(1), 27–37*. Library Literature & Information Science Full Text (H.W. Wilson).
- Montesi, M., & Urdiciain, B. G. (2005). Abstracts: Problems classified from the user perspective. *Journal of Information Science, 31(6), 515–526*. Library Literature & Information Science Full Text (H.W. Wilson).
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE, 6(6), e21101*. <https://doi.org/10.1371/journal.pone.0021101>
- Wang, H.-R. (2016). *How to write a good Data in Brief article*. 5.

Special Thanks

