Evidence Based Library and Information Practice

Evidence Summary

An Online Community of Data Enthusiasts Collaborates to Seek, Share, and Make Sense of Data

A Review of:
Stvilia, B., & Gibradze, L. (2022). Seeking and sharing datasets in an online community of data enthusiasts. *Library & Information Science Research 44*(3). https://doi.org/10.1016/j.lisr.2022.101160

Reviewed by:
Jordan Patterson
Associate Librarian
A. P. Mahoney Library
St. Peter's Seminary
London, Ontario, Canada
Email: jpatte46@uwo.ca

Received: 16 Nov. 2022 **Accepted:** 25 Jan. 2023

DOI: 10.18438/eblip30280

Abstract

Objective – To understand the major activities, tools, sources, and challenges of online communities focused on datasets.

Design – Content analysis informed by activity theory.

Setting – The r/Datasets subreddit, a web forum for sharing, seeking, and discussing datasets.

Subjects – 1232 "hot" or "top" discussion threads (1232 original posts and 6813 responding comments) first posted between 2010 and 2020.

Methods – The researchers used Reddit's API to collect their sample of threads. Using a random subset of the sample, the researchers developed a coding scheme for content analysis, which identified major themes in the data. Through this process, they controlled for quality: each researcher coded half

the subset independently, then together evaluated their intercoder reliability and discussed and resolved disagreements. The researchers also employed labelled latent Dirchlet allocation to construct topic models corresponding to the theme's manual content analysis, which produced profiles of the top 100 terms most likely to appear in that topic. Finally, the researchers extracted URLs from threads in the sample to ascertain types of information and data sources used by the community. Presenting their findings, the researchers discussed notable themes and proposed a metadata model for describing datasets, the Data Q&A metadata (DQAM) model.

Main Results – The r/Datasets community engages in three distinct activities: asking and answering questions, disseminating information, and community building. The closely related Q&A and dissemination activities shared themes of obtaining and aggregating data, sensemaking, collaborating and crowdsourcing, and data evaluation. Community members frequently discussed tools, competencies, and sources for data work. Major challenges for members of the community related to the general themes of data quality, accessibility, ethics, and legality. A proposed 16-element metadata schema should meet the needs of data enthusiasts.

Conclusion – The content analysis reveals a dedicated community engaged in an array of data-seeking and data-sharing activities. Data producers should be mindful of how their data can be accessed and used outside of their original professional or scholarly contexts.

Commentary

Where once datasets were the preserve of ivory-tower statisticians and scholars, the tech competencies of digital natives and a trend toward openness in the Information Age have made data wrangling a viable hobby, to which the existence of the subreddit r/Datasets stands as a testament. The great proliferation of data has been attended by a movement to regulate and systematize the sharing and seeking of data. In Canada, for instance, federal fund-granting agencies CIHR, NSERC, and SSHRC (2021) require researchers to have research data management (RDM) plans in place in order to qualify for grants, hastening the development of data infrastructure online. The researchers ask how data enthusiasts operate in this landscape.

Assessed with Glynn's (2006) critical appraisal tool, this study meets commonly accepted standards of validity. The researchers took evident care in their research design to ensure their content analysis coding scheme was reliable and free from bias; for instance, in developing their coding scheme, the researchers first worked independently to identify themes before coming together to compare their results and generate a final set of thematic categories.

The researchers' DQAM model is sure to be welcomed by casual and professional data enthusiasts alike, but it enters the study abruptly. A short literature review covering the state of dataset metadata schemas would have been a welcome addition. In writing about r/Datasets, the researchers focus on a novel, non-academic context, but the problems they uncover in their study are not so novel—the issues identified by data enthusiasts plague data professionals as well, and have for some time. The researchers propose their own solution to these ills with the DQAM model, but they could have fruitfully considered what other solutions were on offer first; for instance, they could have described how major players in other more typically academic areas such as RDM have approached the issue of dataset metadata.

For one example, the Dataverse Project (Institute for Quantitative Social Science [IQSS], n.d.) "is an open source web application to share, preserve, cite, explore, and analyze research data" that has seen great adoption in the scholarly realm, with Harvard at the forefront. When submitting their datasets to an instance of Dataverse, researchers must input descriptive metadata, and for this purpose the project's documentation includes a dataset citation metadata schema with 78 elements (IQSS, 2019). How does the DQAM model differ from this earlier Dataverse schema? What are the merits and

deficiencies of each compared to the other? What makes DQAM stand out? The need for a metadata model custom-built for the non-academic context may exist, but it seems like an oversight to present the DQAM model without discussing other dataset metadata schemas first.

Looking to future studies, the researchers hope to "expand and triangulate" their findings by interviewing r/Datasets community members. The demographic composition of this community would be a compelling question for further research. There seems to be an implicit assumption in this study that community members are *just* data enthusiasts, and it would be interesting to know how many also operate in a professional or scholarly context. The findings of this study demonstrate that r/Datasets community members are highly capable, technically proficient, and ethically concerned -- hallmarks of a commitment to data beyond the amateur. Ultimately, this study underscores the need for researchers and other data-producers to handle their data with care; there is no telling where their data may end up.

References

Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada (NSERC), & Social Sciences and Humanities Research Council of Canada (SSHRC). (2021, March 15). *Tri-agency research data management policy*. Government of Canada. https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy

Glynn, L. (2006). A critical appraisal tool for library and information research. *Library Hi Tech*, *24*(3), 387-399. https://doi.org/10.1108/07378830610692154

Institute for Quantitative Social Science. (n.d.). *About*. The Dataverse Project. https://dataverse.org/about

Institute for Quantitative Social Science (2019, November 21). *Dataverse documentation v.4.18.1: User guide: Appendix.* The Dataverse Project. https://guides.dataverse.org/en/4.18.1/user/appendix.html

Stvilia, B., & Gibradze, L. (2022). Seeking and sharing datasets in an online community of data enthusiasts. *Library & Information Science Research 44*(3). https://doi.org/10.1016/j.lisr.2022.101160