Investigation of Contexts of Data Discovery and Selection Criteria for Clinical Trials Data

Authors: Ying-Hsang Liu[1], Mingfang Wu[2], Megan Power[3]
Author's Affiliations: [1] Uppsala University, [2] The Australian Research Data Commons, [3] Monash University
Keywords: Data discovery contexts, Data selection criteria, Data reuse, Clinical trials data

STRUCTURED ABSTRACT

Aim of your contribution
This study contributes to the conference theme of processes and practices in information science by investigating the research data discovery contexts of clinical trials researchers who reuse open clinical trials and related health data.

Value of your contribution
We investigated a range of data discovery contexts faced by clinical trials researchers to better understand how data discoverability leads to data reuse. Our findings connect ways researchers both initiate personal data journeys and follow highly managed processes of data selection criteria. We present this to demonstrate the need for a holistic view in understanding data discovery approaches by clinical trials researchers.

Research outline
- Research context
In recent years, the open data movement has grown significantly, leading to increased attention on data discovery and reuse by various stakeholders such as funding agencies, research organizations, data service providers, data practitioners, and end-users. Many countries have responded to the rising global demand for open data by investing in data management and cataloguing, promoting data sharing and citation, and establishing infrastructures to streamline capturing, processing, sharing/publishing, discovering and accessing datasets.

The Australian Research Data Commons (ARDC) is the primary agency responsible for supporting shared research data resources in Australia. The goal of this project is to investigate the various contexts of data discovery that researchers encounter, which is a critical factor that influences data discoverability and leads to data reuse.

- Problem/overall research aims/research question
As part of a larger research project, this presentation focuses on characterising the contexts of data discovery of clinical trials data to inform the design and evaluation of data discovery systems and services. Our research questions are:

1. What criteria do researchers apply for assessing relevance and usability of datasets?
2. What are the relationships between the data discovery contexts and the selection criteria of data search?

- Design/methodology/ research approach
This study has been designed to elicit the contexts of data discovery by adopting a mixed-method approach within the post-positivist research paradigm (Williamson, 2018). Specifically, we used the methods of survey and in-depth interview to understand the broader contexts of data discovery within people's information-seeking processes. A pre-interview survey was administered to learn more about the participant's background, including their research areas/topics, stage of career, job roles and their data sources in recent projects. This information was also used in the follow-up semi-structured in-depth interview. A critical incident technique was used to elicit the contexts of data discovery by asking probing questions (Davenport, 2010; Flanagan 1954). The interview protocol was developed and organised by referencing stages of a data lifecycle adapted from Yong et al (2021).

A total of 17 participants were purposively sampled for the study, sourced directly through the ARDC. The recruitment targeted groups who had prior experience with the HeSANDA (Health Studies Australian National Data Asset) Program or were engaged in clinical trials-related activity, invited through direct invitations sent through the HeSANDA node network. The researchers also actively recruited in certain cases by directly reaching out to recommended or known groups. The participants held varied roles (e.g., a clinical or health guideline developer; a research analyst using clinical trials data; a clinical trials designer or other) and represented a range of career stages. While the majority of participants were engaged with collaborative clinical/health research projects and have extensive experiences discovering and reusing data, we also interviewed clinical trial designers and clinical/health guideline developers who brought multiple perspectives to the study. We also interviewed researchers who we noted engaged more intensively in data discovery to support the tasks of systematic reviews, meta-analysis and guideline development.

- Findings
Our findings suggest that researchers use a diverse set of selection criteria to assess the relevance and usability of datasets depending on the search contexts. We identified shared data attributes in the study and the information considered most important for properly selecting or using a dataset in the European Open Science Cloud (EOSC) data quality attributes (Lacagnina et al., 2023). Moreover, we have examined the relationships between the different contexts of data discovery and their respective selection criteria. Specifically,

we have analyzed the connection between the context of making sense of data and the selection criteria for a user guide, which takes into account factors such as data strengths, limitations, known issues, and the availability of uncertainty information. Additionally, we have explored the relationship between the context of data collection in meta-analysis and the selection criteria for a data dictionary, which requires adherence to community standards for metadata, provision of information on provenance and traceability, and furnishing details about the data provider and point of contact.

- Research limitations/implications
Based on our findings, it is crucial to prioritize the information needs of users when selecting and using datasets. Given that using a dataset entails a multifaceted process of conducting secondary data analysis within a research project, it is essential for data repositories to specify their system design objectives in line with meeting user information needs. One way of achieving this is by enhancing the index and search interface with data variables, such as clinical metrics of outcome measures and methods, thereby making it easier for users to select and use datasets that align with their research needs.

References

Davenport, E. (2010). Confessional methods and everyday life information seeking. *Annual Review of Information Science and Technology*, *44*, 533–562. https://doi.org/10.1002/aris.2010.1440440119

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4): 327–359. https://doi.org/10.1037/h0061470

Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M. E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H., & Dennis, R. (2023). Towards a data quality framework for EOSC. https://doi.org/10.5281/zenodo.7515816

Williamson, K. (2018). Research concepts. In K. Williamson & G. (Eds.), *Research Methods: Information, Systems, and Contexts* (2nd ed., pp. 3–25). Chandos Publishing. https://doi.org/10.1016/B978-0-08-102220-7.00001-7

Yong, A., Zahuranec, A. and Verhulst, S. (2021). A layered approach to documenting how the third wave of open data can provide societal value. Accessed on March 7, 2023 at: https://opendatapolicylab.org/articles/the-onion-model-a-layered-approach-to-documenting-how-the-third-wave-of-open-data-can-provide-societal-value/index.html

Suggested Citation
Liu, Y.-H., Wu, M., & Power, M. (2023). Investigation of contexts of data discovery and selection criteria for clinical trials data. Trends in Information Science – The ASIS&ST European Chapter Research Series. June 20-21, 2023. Virtual conference.