Scholars' Data Reuse Behaviors in Disciplinary Context: A Meta-Synthesis Study

Xiaoguang Wang[1][0000-0003-1284-7164], Qingyu Duan[2][0000-0002-6881-1825], and Mengli Liang[3][0000-0003-1160-9957]

[1] Wuhan University, CN wxguang@whu.edu.cn
[2] Wuhan University, CN Duanqingyu@whu.edu.cn
[3] Wuhan University, CN liangmengli@whu.edu.cn

Abstract. Data reuse plays a pivotal role in science research in the data era. Given that the impact of discipline culture on data reuse is deeply rooted, we explore data reuse behaviors of the two groups of scholars with significantly different qualities, the nature science and the humanities and social science. Relying on the meta-synthesis and inductive coding approach, information about intentions, influence factors, data processing and using and data reuse barriers were extracted from 37 qualified articles and then analyzed. Results show: 1) informal channels perform a vital role in data reuse in both two communities; 2) there is a distinct correlation between data reuse and disciplinary context. 3) clear distinctions exist between two fields in data reuse barriers, disciplinary practice degrees and data reuse patterns. The results imply the urgency to establish data managers, link publications and data, and enhance data organization.

Keywords: Data Reuse Behavior, Researchers, Field Comparison, Meta-Synthesis.

1 Introduction

With the rapid growth of data and the development of data-driven science paradigm, countries and academic communities are preparing for data reuse focusing on sharing and management, so as to dig up the additional value of primary data. Data reuse is rarely formally defined, but is generally considered to repurpose old data for a new problem（Zimmerman，2008）. In recent years, the literatures on data reuse have mushroomed, which have shown that data reuse behaviors vary across disciplines (Tenopir, 2015), driven by their respective resource systems, data infrastructure and research paradigm (Borgman, 2008). However, the commonalities and differences in data reuse across disciplines are rarely discussed in detail, though many studies have explored the researchers' data reuse behaviors in specific disciplines (Joo et al., 2017). In this poster, we conducted a meta-synthesis study to explore the relationship between data reuse and disciplinary context, which is closely related to the improvement of data management and data service.

2 Methods

We applied meta-synthesis method for the study, which was aimed to analyze by synthesis those literatures with the empirical approach and the topic of scholars' data reuse behavior. Firstly, all relevant literatures were retrieved in the Web of Science core collection and ProQuest with "data reuse" and "secondary data analysis". 20 articles were selected after excluding irrelevant and non-empirical studies. Secondly, based on the original pool of 20, snowballing-based citation tracking and reference tracking were conducted to discover more potential articles, where literatures about engineering or without clear discipline descriptions were removed from further consideration, and 59 studies were captured finally. Thirdly, according to the EBL critical evaluation in Library and Information, Glynn's checklist(Glynn, 2006) and Israel's sample size checklist(Israel, 2009) were chose as measure tools to identify the literatures with a score of more than 75%, in which sample features, data collection procedures, study design and study results of each study were quantitatively evaluated.

For each qualified article, we abstracted the research topic, disciplinary features and basic descriptive information about procedures, intentions, influence factors, sources, methods, types, barriers of data reuse. Through iterative review, we derived the essential categories of scholars' data reuse behaviors with inductive coding approach.

3 Results and Discussion

A total of 37 articles were included in the follow-up analysis, 14 of which were about natural science scholars, 14 about humanities and social science scholars, and 8 involves multi-community comparison, covering various disciplines, such as life science, health science, ecology, biomedical science, clinical science, astronomy, physics, biology, anthropology, politics, library science, archaeology, psychology, art, history, and linguistics.

3.1 Data Reuse Ecosystem

The study finds there are three categories in data reuse ecosystem, namely intentions, influence factors and reusing process. Especially, we measured the frequency of assistants in the reuse process, following the principle: if it appears in the study for successful reuse, it will be recorded once; if it ranks 1 in that survey, it will be recorded twice. The statistical result is described as Figure 1 and data reuse ecosystem is depicted as Figure 2.

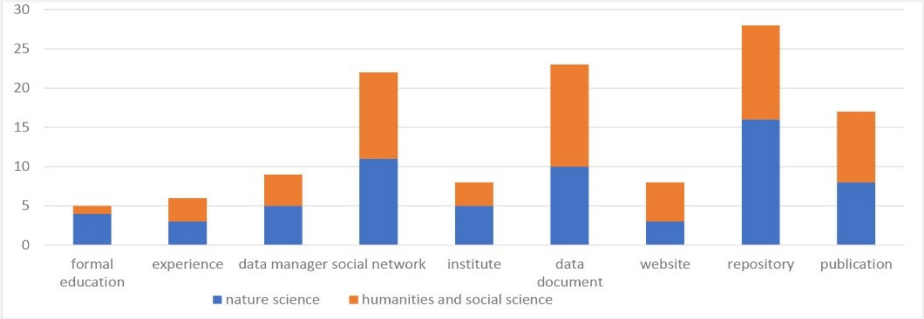

Fig. 1. The frequency of assistant in reuse process

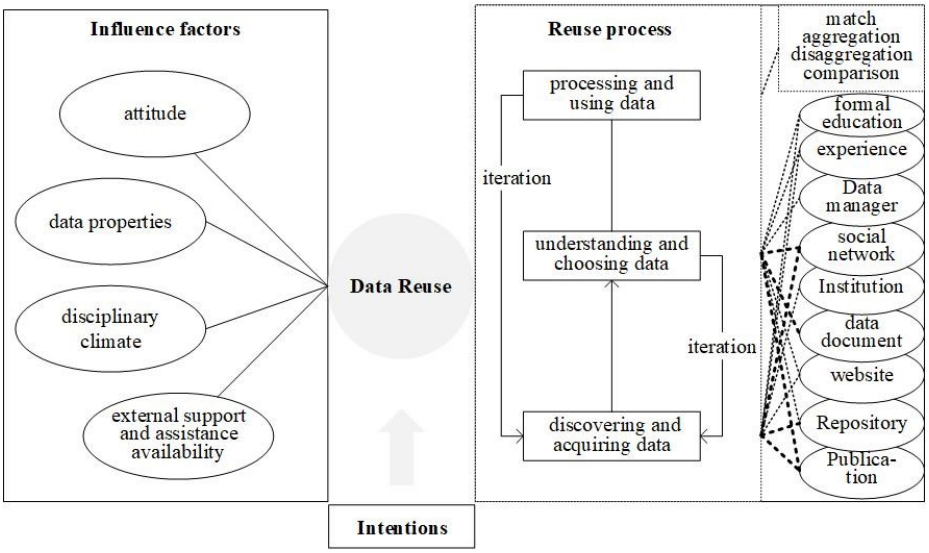

Fig. 2. Data reuse ecosystem

In the below, the data ecosystem is described in detail from three categories:

1) Intentions to data reuse. The study finds scholars reuse existing data in nature science for 7 goals, including original study, meta-analysis study, model/theory/method test, software or tool development, datasets comparison and experiment control, reproducibility study and infrastructure development(Federer, 2019). However, the intentions of scholars in humanities and social science are rarely discussed comprehensively, with only sporadic references to large sample data, education and training at lower research costs.

2) Influence factors to data reuse. There are 4 types of factors influencing scholars' behaviors, details of which are similar in the two communities, namely scholar's attitude to reuse, data properties, disciplinary climate and external support and assistance

availability. Perceived usefulness and perceived concern will produce an effect on scholars' attitude between two communities, but perceived effort is another significant factors in humanities and social science, referring to the time and energy (Yoon et al, 2017). Data properties are mostly composed of relevance, availability, accessibility, understandability, quality, reputation of data collector (Faniel et al, 2016). Disciplinary climate involves norms of data reuse, data openness, disciplinary receptiveness and peer encouragement. External support and assistance availability are affected by repositories availability, primary collector reach, human intermediaries reach, training and expertise, institutions' support (library, museum etc.).

3) Reuse process. The process is iterative and nonlinear, consisting of three steps, where social network (colleagues, mentors, primary collector etc.) in reality plays an important role for discovering, acquiring, and understanding data with e-mail, discussion, visit or direct request, although both formal channels and informal channels are contribute to scholars' data reuse (see Fig. 1). In addition, publications are considered the most common way for discovering the clue whether data exist and where they are (Zimmerman, 2007). Furthermore, nature science scholars tend to acquire data from repositories, while the humanities and social sciences scholars also often get their data from official agencies' websites (see Fig. 1).

When processing and using data, aggregating multi-source data is the basic approach for comparative analysis, context reconstruction, annotation, presentation, etc., because it is generally recognized that datasets are more valuable than a dataset. Moreover, other typical approaches are disaggregation, comparison, and matching. (Whitmore, 2016).

3.2 Field Comparison

The study also shows that 9 barriers exist in data reuse process (see Table 1) and their prevalence is calculated as figure 4, following if the primary concept occurs in the article, the frequency of its category will be increased by 1, even if it appears several times. As figure 4 presents, the most universal barrier is data openness, followed by sparse data documents, poor data quality, limited human scaffolding, poor data interoperability and concern risks. In terms of data openness, the research data collections at individual levels are lower than resource or community data collections and reference collections, resulting from data stored locally.

Table 1. Results of inductive coding

Primary Concepts	Categories
Ethical consideration, confidentiality, personal data stored locally, low digitization, limited accessibility, not free, exclusive and unfamiliar data software or analytics programs, a lack of reward for sharing, ownership issues, lacking data citation practice, not recognized as an academic achievement	Low data openness
A lack of programmatic documents, unsuitable order of metadata records	Sparse data documents
Poor accuracy, insufficient integrity, data errors, improper cleaning, lacking organization, sample selection bias	Poor data quality
Diversity data formats， metadata and documentations schemes, incompatible Levels of granularity, least-developed standards, multi-sources following different conceptual standards, a low level of data integration	Poor data interoperability
Limited publishing room, original value being questioned	Low status
---	Primary data dispersing
Available bandwidth and CPU space limit	Technical issues
Primary collector reach failure, lacking awareness of data manager and information experts existing	Limited human scaffolding
Fear of data abuse/misinterpret concern, fear of infringing ethical codes, fear of data errors, fear of copyright infringement	Concern risks

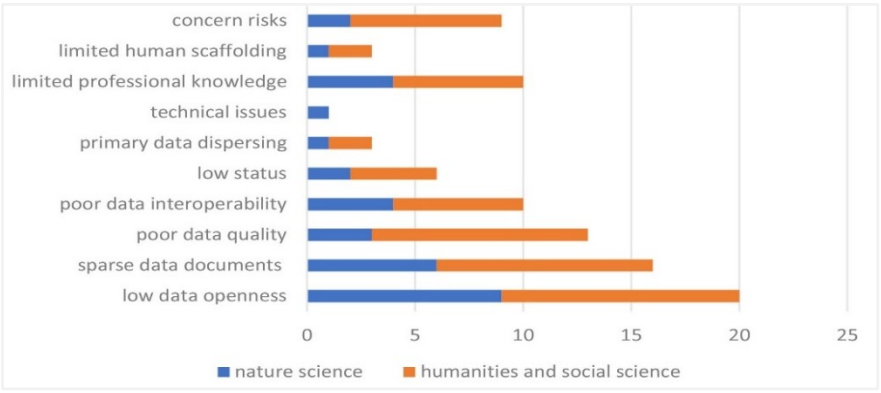

Fig. 3. The frequency of barriers in data reuse

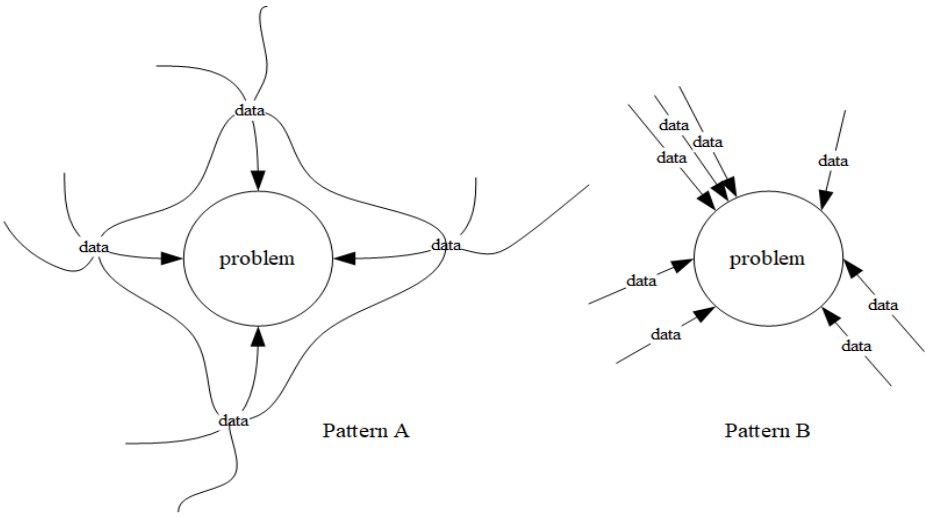

Fig. 4. The patterns of data reuse

Furthermore, the finds from the study make it clear that how and why data reuse behaviors vary in different communities:

1) Disciplinary receptiveness and maturity: scholars in nature science face fewer barriers than in humanities and social science, especially on concern risks and poor data quality9 (see Fig. 3). This means disciplinary climate of nature science about data reuse is more superior to that of the humanities and social science, which are still skeptical of data reuse and have poor infrastructure.

2)Focus of barriers: a lack of programmatic documents are considered a crucial barrier for nature science scholars while humanities and social science scholars particularly concern with a lack of descriptive information in metadata. Because the collection information of natural science data can be easily copied into the tag, while the context information of humanities and social science data tend to scatter in a series of data independently stored, like documents, pictures, filed notes, videos, which need metadata records to determine the relationship with each other data object and the complete context (Daniels，2014).

3) Structure of reuse patterns: There are two patterns: the pattern of net centripetal structure (pattern A) and the pattern of linear centripetal structure (pattern B) (see Fig. 4). On the whole, nature science scientists use the pattern A more frequently, while humanities and social sciences researchers tend to use the pattern B. This is caused by two factors. One is the shift from print to digital in humanities and social science but data in the natural science have long been digitized. The other is the data organization superiority in nature science, leading to some data alliance, in which relevant data can be mapped to each other and automatic aggregated. Nevertheless, the most data of humanities and social sciences are acquired through data island and then aggregated by scholars. Of course, here is crossover and dual use of the patterns among the nature science scholars and the humanities and social science scholars.

4 Conclusion

Data reuse reflects the state of the data infrastructure, which is closely related to disciplinary progress. In this poster, we reported some preliminary results about scholars' data reuse behaviors in disciplinary context. We summarized the data reuse ecosystem, where informal channels are considered significant sources and assistants, and publications and repositories are regard as pivotal formal channels by all scholars. However, it was proved that scholars' data reuse behaviors are different in two communities, typically in the barriers, the disciplinary practice degrees and the patterns. In addition, the 9 barriers of data reuse shown in the poster imply the urgency of the establishment of professional data managers for institutions, the association between data and publications and the enhancement of data organization.

The poster is a theoretical and secondary analysis result. In the future, we will concern with the further demonstration of discipline patterns of data reuse, and exploring how publications can be associated with data from the linked open data and semantic publishing perspectives.

References

1. Zimmerman, A.: New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values 33(5), 631-652 (2008).
2. Tenopir, C., Allard, S., Douglass, K., et al.: Data sharing by scientists: practices and perceptions. PloS one 6(6), e21101 (2011).
3. Borgman, C. L.: Data, disciplines, and scholarly publishing. Learned Publishing 21(1), 29-38 (2008).
4. Joo, S., Kim, S., Kim Y.: An exploratory study of health scientists' data reuse behaviors: Examining attitudinal, social, and resource factors. Aslib Journal of Information Management 69(4), 389-407(2017).
5. Glynn, L. A.: Critical appraisal tool for library and information researc. Library Hi Tech 24(3), 387-399 (2006).
6. Israel, G. D.: Determine sample size (2009), https://www.tarleton.edu/academicassessment/documents/Samplesize.pdf, last accessed 2019/9/23.
7. Federer, L.: Who, what, when, where, and why? quantifying and understanding biomedical data reuse. University of Maryland, College Park, Maryland, USA (2019).
8. Yoon, A., Kim, Y.: Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. Library & Information Science Research, 39(3), 224-233 (2017).
9. Faniel, I. M., Kriesberg, A., Yakel, E.: Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology 67(6), 1404-1416(2016).
10. Zimmerman A.: Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries 7(1-2): 5-16(2007).
11. Whitmore, D. A.: Seeking Context: Archaeological Practices Surrounding the Reuse of Spatial Information. University OF California, Los Angeles (2016).
12. Daniels, M. G.: Data Reuse in Museum Contexts: Experiences of Archaeologists and Botanists. University of Michigan, Michigan (2014).