Toward Data Quality Assurance Infrastructure for Research Data Repositories

Project Summary

Florida State University (FSU), the lead applicant, in partnership with Texas A&M University (TAMU), is requesting $92,995 of IMLS funds to complete a collaborative one-year research project. The project will contribute towards the IMLS NLG program Goal 3 and Objective 3.2: “Support innovative approaches to digital collection management.” This exploratory research project will identify and inventory the practices of data quality assurance (DQA) of research data repositories. In particular, the study will identify the types of data quality problems and incidents; models, standards, and strategies used; the division of labor and the roles played; challenges and barriers to evaluating and maintaining data quality, and skills and competencies needed. In addition, the project will use the study's findings to develop DQA design scenarios. DQA scenarios will describe specific data quality assurance problems or incidents, and related DQA actions. We will share the study's findings, including scenarios, with library and academic communities. The study's findings can inform data curators' DQA work, help them identify data quality problems and address those problems.

Project Justification

Quality is defined as "fitness for use" [13].#***Data quality, along with privacy and access, are critical ethical aspects of data use. In the era of Big Data and a flood of research data and publications, the old dictum "garbage in garbage out" is as relevant as ever. Quality of data determines the quality of research findings, teaching, business decisions, policies, and ultimately, it affects human lives***#[17,22]. Universities are presently making significant investments in developing trustworthy and secure infrastructure to curate digital research datasets produced and/or used by their faculty and students. These efforts are motivated by faculty's need to preserve and share their research data [19,25]; state and federal funding agencies requiring their grantees to share data with open access to benefit taxpayers, to enable its reuse in research and teaching, and to enhance research reproducibility and replicability [18,19]; national and state laws that require ensuring the quality of data and preserving individuals' privacy [3,27], and universities' desire to enhance their visibility and reputation by providing open access to data produced by their faculty and students as public goods to benefit their states and society in general [26]. ***One of the main inhibitors of data sharing and reuse, however, is concern about the quality of data.*** Data owners can be concerned about the quality and/or documentation of their data and its potential misuse or misinterpretation by others [23]. The users, on the other hand, need useful, valid, and trustworthy data that represents the phenomena they are interested in, not just Big Data [4,20]. They may not have access to and/or knowledge of the process that generated and/or manipulated the data needed to evaluate its quality. Hence, they may mistrust and not use the data. Furthermore, data is often incomplete and contains biases and inaccuracies. Data creators usually collect or assemble datasets for specific purposes or uses. If data is not properly documented, understanding those purposes is often a challenge and a barrier to data reuse [24]. Furthermore, DQA is not free. Digital data repositories need to find cost-effective, efficient ways to evaluate, maintain, and communicate the quality of the data they share to facilitate its ethical reuse. There have been conceptualizations of research data quality and studies of scientist perceptions and priorities for data quality and data quality assurance skills [e.g., 7,10,12,23]. The perception of what constitutes quality and useful data and/or when the data becomes useful may vary within the same process, discipline, and across different processes within disciplines [11]. Scientists may rely on different properties and cues of data to assess its relevance, value, and reusability [8,23]. There have been several general quality assurance standards and approaches used in the industry (e.g., Six Sigma, ISO 9000). Likewise, there is a significant body of literature on and a few models of data curation [e.g., 2,5,11,15,16]. There is, however, ***a lack of in-depth empirical studies that focus on DQA practices and activities in research data repositories***. DQA activities may range from quality evaluation and improvement actions performed by data providers and repository staff to data cleaning performed by students as part of their class assignment, or DQA hackathons and research reproducibility challenges[1]. This study will address this gap by examining the following research questions:

1. What data quality problems, challenges, and incidents do data curators/repository managers encounter in their work, and how do they resolve those problems?

[1] https://paperswithcode.com/rc2021

a. What DQA methodologies, standards, workflows, metrics, and metadata do they use to evaluate and ensure the quality of datasets they curate and to communicate that quality level/status to users?
b. How is the DQA work divided? What roles are played? Does the DQA practice involve the original contributors and reusers of data? Who is responsible for the quality of reused data, including ethical issues that may stem from the reuse of low quality data?
c. What skills and competencies are needed for successful DQA?

Project Work Plan

This research study will be guided by a theoretical framework that comprises activity theory [14], information quality and information credibility frameworks [6,22], and self-determination theory [21]. We will use activity theory to conceptualize general structures of DQA activities and problems. The information quality and credibility evaluation frameworks will help us model the structure of data quality and credibility evaluation and relations among data use activities and data quality problems. Finally, we will use self-determination theory to develop interview protocol questions related to motivations for contributing to DQA.

The study will begin with an analysis of documentary evidence. We will sample 50 data repositories from the re3data.org registry that are run by research universities or university consortiums/systems. We will analyze repositories' data curation service descriptions, policies, data use/reuse agreements, and metadata schemas and vocabularies. Findings of the documentary analysis will be used in selecting interview participants and developing an interview protocol. Next, we will conduct semi-structured interviews with 30 data repository managers and curators. We will use thematic content analysis to analyze documentary and interview data. Finally, we will apply scenario-based task analysis [9] to the findings of the documentary and interview data analysis to develop DQA design scenarios and recommendations. Potential participants will be identified from an analysis of the repository websites, as well as by using a snowball sampling approach.

Dr. Besiki Stvilia, Professor in the School of Information at FSU, will serve as the Project PI and Director. He brings to the project expertise in the areas of data and information quality and data curation. Dr. Stvilia will lead the overall effort to conduct the proposed research and assemble DQA scenarios and recommendations, supported by a doctoral research assistant (RA). Dr. Dong Joon (D.J.) Lee, a TAMU Associate Professor and Research Information Systems Librarian, will serve as the Project Co-PI. He brings to the project expertise in research information management and research data curation. He will lead on the content analysis of documentary data collected from research data repository portals.

Project Results

This applied research project will inform the design and construction of digital data curation infrastructure components on university campuses that aim to provide access not just to Big Data but reusable, trustworthy data - data that could be used with confidence in research, teaching, policymaking, and developing services for consumers and society in general. The study's findings, DQA scenarios and recommendations will inform data curators' DQA practices, including the design of their services, data stewardship metadata, and policies. The study's outcomes will also inform the data curation and data science training and education curricula of LIS schools and communities of practice. We will distribute research findings of this project at two library conferences (e.g., ACRL 2023, IDCC 2023) via presentations and peer-reviewed journal publications. Publications, presentations, and data, including DQA scenarios and recommendations, will be posted on the project website, deposited at the FSU and TAMU data repositories, and publicized in research data curation and research data management communities through community listservs.

Budget Summary

The total requested amount from IMLS is $92,995. This includes one month of summer salary for Dr. Stvilia ($15,655), two semesters of salary and tuition for a graduate RA ($25,247), and their travel to one conference ($5,000); incentives for participants who consent to be interviewed ($30 x 30 = $900). The TAMU subaward budget includes one month of summer salary for Dr. Lee and his travel to one conference ($16,183). F&A will be charged according to FSU's federally negotiated rates (54% until 06/30/23) on MTDC for a total of $55,575. FSU commits $59,186 as cost share in the form of 15% of Stvilia's time for Fall 2022 and Spring 2023.