Beyond Visuals: Improving Accessibility of Data Curation and Multi-Modal Representations for People of all Abilities through Reproducible Workflows

The School of Information Sciences at the University of Illinois at Urbana-Champaign (UIUC) requests $532,510 to support Dr. JooYoung Seo's three-year Laura Bush 21st Century Librarian Early-Career Development project that addresses program *Goal 3*, and *Objective 3.2* and *3.4*. In partnership with Posit Public Benefit Corporation (formally, RStudio), the Chart2Music (C2M) open-source project team, the Data Curation Network (DCN), and the National Federation of the Blind (NFB), the PI will address imperative needs of developing accessible data curation and inclusive data visualization tools for both professional data curators and participants of all abilities, including people with visual impairments. The project results will lead to the development of accessible open-source tools that can be integrated into reproducible research workflows by which data curators can produce more inclusive data visualizations for people of all abilities.

Project Justification: With the advent of computational reproducibility frameworks (e.g., Jupyter Notebook; R Markdown), reusable and transparent data curation has become easier than ever before among data librarians and curators. Furthermore, the recent development of the open-source scientific publishing system, Quarto, has opened another door for researchers and data curators to share the whole process of collecting, managing, discovering, and presenting reusable data with the general public through its language-agnostic literate programming interface. Despite the growing benefits of such reproducible data curation frameworks, some groups of people still remain marginalized from public access to the curated data due to the following fundamental accessibility barriers: (1) Data curation is often carried out without deep understanding of accessibility; (2) Data curators lack practical training and useful tools that can significantly improve the accessibility of their data curation (beyond simple alt text). For example, two widely adopted data representation methods by data curators are numeric tables and visual graphs. Without careful consideration, these formats may not be readily accessible to people with visual impairments (e.g., blindness; low-vision; color-blindness). However, it is quite challenging for day-to-day data curators to come up with inclusive solutions that can make materials accessible. The best effort that accessibility-mindful curators can make is merely to conform to the Web Content Accessibility Guidelines, such as adding alternative text to images and semantic tags (e.g., headings and landmarks). This is never an ideal solution especially for data curation and visualization because such retrofitted accessibility patches often remain an extra step for the curators' judgment call outside of the reproducible workflows. Moreover, the quality of the accessible curations heavily depends on the data curators' prior experience with accessibility. To tackle this issue more effectively and to come up with a more consistent and robust solution, this project aims to develop toolkits that bring accessibility to existing reproducible data curation workflows as part of its pivotal components.

Project Work Plan: This three-year Early Career Development project addresses three research questions: (1) How can we better support data curators to easily improve the accessibility of curated data (e.g., data charting and visualization) for people with a wide range of sensory, physical, and cognitive abilities? (2) In what ways can accessibility be integrated into existing computational tools and reproducible frameworks as part of essential components? (3) What design considerations contribute to inclusive and multi-modal data representations going beyond a single visual modality? To address these questions, the project is planned as follows.

In Year 1, in close partnership with Posit PBC and the C2M open-source team, the PI will develop a cross-language data science package across R and Python that can translate each of the most widely used visualization objects (i.e., R ggplot; Python matplotlib) into accessible multi-modal (e.g., audible; touchable; readable) representations. Since each visualization package has their own graphic data structure, the PI and a software engineer will first design the abstract syntax tree model that can programmatically capture the visual layers (e.g., graph type; axis values and labels; aesthetic mappings and geometries) from ggplot and matplotlib objects, and will map them to a language-agnostic API that can augment these metadata into multi-modal formats. In sonification mode, for instance, stereo panning sound can represent x-axis from left to right; different tones can correspond to y-axis values. In tactile mode, Braille Unicode can be used to

represent the overall data patterns (e.g., ⠤⠤⠒⠒⠉⠉) that can be read through end-users' refreshable braille displays. As this multi-modal data representation package can be integrated into many existing computational reproducibility systems (e.g., R Markdown, Jupyter Notebook, and Quarto), it is also possible to auto-generate text-based chart summaries and descriptions through the interface communications between the computational back-end engine and the language-agnostic front-end API. The core research and development will be guided by the PI's related expertise. The PI as an award-winning blind information scientist (*LG-252360-OLS-22*) and emerging learning science scholar recognized by the *International Society of the Learning Sciences* (ISLS) in 2022, he has been leading the ISLS-sponsored "Data Accessibilization" project at the University of Illinois, and contributed to the accessibility enhancements of multiple open-source data science packages through his engineering skills on GitHub (e.g., rmarkdown; knitr; bookdown; shiny; gt; distill; rticles; quarto). In addition to the PI's expertise, the close partnership with Posit PBC and C2M open-source team will provide necessary technical support for this project (e.g., ensuring that accessibility API updates in their products are compatible with the PI's toolkits).

In Year 2, with approved IRB study protocols, the developed toolkits will be evaluated and refined through iterative feedback from two end-user groups: (1) professional data curators from DCN; (2) blind and low-vision users from NFB. The PI and his RA will start with pilot user studies with 5-10 experts from each group to reflect both curators' and patrons' real experiences with our system. Data curation experts will inform us of the usability and usefulness of the toolkits in turning their data into accessible multi-modal representations. We will also observe how our tool can be unobtrusively integrated into data curators' reproducible workflows. Blind and low-vision users will test the accessibility and understandability of the multi-modal data representations curated by the data professionals. Our team will examine how blind users customize different modalities for various chart and table types in conjunction with their assistive technologies. The practical feedback collected from both groups will guide us to better system design of our tool. We will keep conducting iterative user studies online using a convenience snowball sampling for both data curators and end-users with disabilities (up to 100 participants), by which we will upgrade our system accordingly.

In Year 3, project outputs will be widely disseminated to the public. The developed package will be released on GitHub as an open-source project with detailed user manuals and online sample galleries. Training and workshops will be provided through data professional networks, including the Posit (formerly RStudio) annual conference, DCN events, and the Champaign-Urbana Data Science User Group. The PI will also publish and present the work results at nationwide and international information conferences, including iConference, ACM ASSETS and CHI, CSUN Assistive Technology Conference, and the NFB Annual Meeting.

Diversity Plan: This project actively addresses diversity and inclusion with the aim to narrow the information access gaps between people with and without dis/abilities. Our team consists of people with diverse backgrounds and dis/abilities, including the PI who is blind and who has an intercultural background. All of the technical partners (Posit and C2M) prioritize diversity and inclusion in their open-source products.

Project Results: The project results will not only remain in the free open-source package development and online training resources for accessible data curation, but also foster a more inclusive culture among professional data curators by engaging them with serving their marginalized community (e.g., blind population). Furthermore, the project results can lead to new curriculum development on accessible data curation at LIS community as a sharable module.

Budget: The proposed budget of $532,510 includes 1 summer month salary in Y1-3 for the PI ($40,677); a full-time software engineer in Y1-2 for system development and maintenance ($98,515); one PhD Research Assistant in Y1-3 ($80,546) with tuition remission ($51,549); fringe benefits ($66,816). Other direct costs include travel to conferences ($12,000); materials and supplies ($2,700); human subject research costs in Y2 ($2,000). Indirect costs are $177,707.

PI Information: Dr. JooYoung Seo is an assistant professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign, internationally certified accessibility professional, and one of the few blind information scientists in the world. His teaching and research involve accessible computing and inclusive data science for people with and without dis/abilities. More information can be found at https://jooyoungseo.ischool.illinois.edu