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

1. Project Justification

The School of Information Sciences at the University of Illinois at Urbana-Champaign (UIUC) requests $649,921 to support Dr. JooYoung Seo's three-year Laura Bush 21st Century Librarian Early-Career Development project that will be conducted from August 2023 through July 2026. This project addresses the Laura Bush 21st Century Librarian Program *Goal 3: Enhance the training and professional development of the library and archival workforce to meet the needs of their communities*, and focuses on *Objective 3.2: Create and/or refine training programs that build library and archival workforce skills and expertise in contributing to the well-being of communities* and *Objective 3.4: Support training of the library and archival workforce to advance digital inclusion for the benefit of community members*. In partnership with the national Center for Supercomputing Applications (NCSA), Posit Public Benefit Corporation (FKA, 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.

1.1 Accessibility Gap in Reproducible Data Curation

In recent years, reproducible frameworks using open-source data science languages (e.g., Jupyter Notebook; R Markdown) have gained popularity for academic research and among data curators and archive professionals (Sawchuk & Khair, 2021). Reproducible frameworks are defined as a standardized method of conducting scientific research using code scripts that produce accurate, transparent and reproducible results (Fekete & Freire, 2020; National Academies of Sciences, 2019). Python Jupyter Notebook (ipynb) is an open-source web application that allows users to code collaboratively with a combination of codes and text explanations (Kluyver et al., 2016; Perez & Granger, 2007). Similarly, R Markdown (Rmd) provides an integrated solution for creating dynamic documents that combine R codes, visualizations and narrative text formats (Allaire et al., 2023; Xie et al., 2018). Both frameworks provide a user-friendly environment to integrate code, documentation, and visualization, offering transparency and reproducibility in academic research and professional data curation. In addition, these frameworks can generate output formats such as HTML, PDF, Markdown, Latex that can be shared across multiple platforms, making them more convenient for disseminating research findings. Furthermore, the recent development of the open-source scientific publishing system, Quarto (Allaire et al., 2022), 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. Literate programming is a programming practice that aims at writing computer programs in such a way that non-experts can understand them easily. It involves dividing code into logical paragraphs and including the necessary documents within the source code to promote clarity and readability, rather than keeping it in separate files or comments. (Knuth, 1984; Leisch, 2002). This approach allows data curators and archive professionals to more easily conceptualize and reason about their code, as well as facilitating collaboration and knowledge sharing among team members (Sawchuk & Khair, 2021).

Despite the growing benefits of such reproducible and literate programming data curation frameworks, some groups of people remain marginalized from public access to the curated data (Joyner et al., 2022; N. W. Kim et al., 2021; Lee et al., 2020; Marriott et al., 2021) This is 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 print disabilities (N. W. Kim et al., 2021; Lee et al., 2020),

including visual impairments (e.g., blindness; low-vision; color-blindness) and cognitive impairments (e.g., Dyslexia; Dyscalculia; Dysgraphia). 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 (see Project Work Plan for details).

1.2 Comparing Current Approaches with PI's Innovative Contribution

There are three typical alternative approaches to data visualization for the blind: (1) data tactilization representing data patterns through haptic devices (Brown & Hurst, 2012; Paneels & Roberts, 2010) or refreshable Braille display (O'Modhrain et al., 2015), (2) data verbalization representing data via text-based natural language descriptions(D. H. Kim et al., 2020; Lundgard & Satyanarayan, 2022; Srinivasan et al., 2021), and (3) data sonification using spatial sound to represent data points (Summers et al., 2019; Vines et al., 2019; Zanella et al., 2022). Unlike sighted people who can utilize multiple senses and strategies to interpret different types of data representations, blind people are often limited to one or two senses of delivery, which calls for the need for an integrated approach. To address this gap, PI Seo and his research team have recently created a proof-of-concept technology that integrates these three approaches, laying the groundwork for further innovative research proposed in this project. The PI has named this front-end interface as MAIDR (Multimodal Access and Interactive Data Representation; https://bit.ly/3YI5n8c). The MAIDR system provides blind users with three customizable multimodality (i.e., braille, text, and sonification: BTS) modes and each mode is togglable via single-letter BTS hotkeys. This interface is fully compatible with assistive technologies (e.g., screen readers; refreshable braille displays) and supports any modern web browsers, including Chrome, Edge, Firefox, and Safari on various operating systems.

PI's scientific contribution and his MAIDR system were featured in Nature as a promising tool that enables visually impaired scientists to interact with scientific data (Katsnelson, 2023), highlighting potential future opportunities. 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, the PI 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; see the letter of commitment from Posit PBC for this evidence). Furthermore, the PI has been dedicated to integrating accessibility research into his teaching in introductory data science courses for his students to learn multiple ways of data representation for a wider audience. The PI has shared this experience in the Journal of Data Science to promote the importance of teaching accessible visualization among other data science instructors (Seo & Dogucu, in press).

The project proposed here extends the PI's previous work and expertise regarding reproducible frameworks and multimodal representation. Not only does this research proposal align well with PI Seo's prior work, but it also contributes significantly to the field of Library and Information Sciences by creating an accessible and inclusive framework. This new knowledge on data visualization and access benefits both visually-impaired individuals and the data curator community.

1.3 Target Population and Beneficiaries

This project aims to (1) develop accessible data visualization tools that can seamlessly be integrated into data curators' reproducible frameworks, such as Python Jupyter Notebook and R Markdown; and (2) make it easier for blind patrons to

comprehend via multimodal data representations. Thus, this project targets two groups: data curators with little to no accessibility training and blind patrons who face challenges in visually curated data.

1.3.1 Data curators and archive professionals

Data curators are responsible for managing scientific research data and ensuring that it is discoverable, accessible, and useful. Data curators need to be aware of best practices in data management and accessibility standards, including web content accessibility guidelines (WCAG 2.1). However, many data curators lack accessibility training and may not be aware of the barriers faced by individuals with disabilities when accessing digital information. This project targets this population by providing easy-to-use tools, accessibility training and resources for accessible data curation.

There are varying estimates on the number of data curators in the United States. According to recent reports, Zippia's 2021 demographics and statistics suggest that over 8,254 individuals are currently employed as curators (Zippia, 2021). Meanwhile, a 2020 workforce report by Data USA estimates up to 61,533 people working as archivists, curators, and museum technicians in the U.S. (Data USA, 2020), while a 2021 Bureau of Labor Statistics report cites over 11,000 employees in these fields (U.S. Bureau of Labor Statistics, 2021). Although it is difficult to obtain an exact estimate of the number of data curators who may benefit from this project, it is clear that a significant number of data professionals can benefit from this project. The integration of new tools and techniques within already-established reproducible workflows will certainly reduce overhead and increase efficiency, thus making data curation more accessible and cost-effective for all beneficiaries of our program.

1.3.2 Blind patrons

The second target population is blind patrons who face challenges in accessing digital information, especially data visualization. Accessible data visualization is essential for blind individuals to understand complex information, such as scientific research data. Unfortunately, many data visualization tools lack accessibility features, making them inaccessible to blind individuals. Therefore, this project aims to enable blind patrons to access and understand curated data through multimodal data representations (see Phase 1 in Project Work Plan).

According to statistics, approximately 40 million individuals worldwide suffer from complete blindness or moderate to severe vision impairments (Bourne et al., 2021). In the United States, around 7 million people, which accounts for 2.17% of the population, were living with irreversible visual acuity loss or blindness in 2017 (Flaxman et al., 2021). However, there is no universally accepted definition for blindness, and there exist a wide range of general and specific variations describing the level of blindness, including visual disabilities, visual impairments, partial sight, low vision, and vision loss. Therefore, it is crucial to describe the scope of blindness for this project. In this proposal, we refer to blind people, resting upon the following two guidelines: (1) as a statutory definition in the US, “legally blind” is that central visual acuity must be 20/200 or less in the better eye with the best possible correction or that the visual field must be twenty degrees or less (Social Security Administration, n.d.); (2) as a functional and sociological definition suggested by the NFB, one is blind to the extent that they must devise alternative techniques to efficiently do those things which they would do if they had normal vision (National Federation of the Blind, n.d.). Our target audience includes individuals who may not be completely blind, but experience functional low vision or severe vision impairments.

It's worth mentioning that in this proposal, we adopt identity-first language when referring to our blind target population, such as "blind people," rather than person-first language, which uses phrases like "people with visual impairments." Although person-first language is a common recommendation in various academic fields, some disability communities prefer identity-first language (Sharif et al., 2022). Since both our community partner, NFB, and the blind PI of this proposal consider blindness as an identifying factor rather than a negative characteristic, we will use identity-first language throughout this proposal and beyond. Nevertheless, this is merely our preference for wording and does not necessarily reflect the language preferences of each individual who is blind.

Given the aforementioned definitions, our target population consists of blind individuals interested in accessing data archives, as well as prospective blind patrons who can benefit from our proposed accessible data visualization system.

1.3.3 Underrepresented blind scientists and researchers

This proposed project will also carry the potential to inspire more blind researchers and scientists, who like PI Seo, may face barriers due to accessibility issues within their respective fields. While there is a lack of comprehensive data on the number of blind scientists currently working in research institutions, a recent study found that out of 52,124 researchers applying for funding from the US National Institutes of Health (NIH) in 2018, less than 100 self-identified as having a visual impairment, indicating a serious underrepresentation of visually impaired scientists in the field (Katsnelson, 2023; Swenor et al., 2020). The successful implementation of this project can set an example for other funders to create new grant opportunities for blind scientists, to design and test further innovative accessible technologies, thereby facilitating the promotion of new knowledge among different communities that have been previously excluded.

2. Project Work Plan

2.1 Theoretical Background

This project builds upon two theoretical backgrounds and attempts to propose a new way to integrate them towards a more accessible data curation: (1) multimodality (Kress, 2010) and (2) cognitive theory of multimedia learning (Mayer, 2014). Multimodal theory posits that people communicate and express themselves through many different modes. A mode is generally defined as a communication channel that a culture recognizes, including writing, gestures, posture, gaze, font choice and color, images, videos, and even the interactions between them (Kress, 2010). With the technological advances in multimedia, cognitive scientists have highlighted the importance of multimodal communication in human learning and information process. For example, Mayer (2014), in his cognitive theory of multimedia learning (CTML), suggests effective instructional design strategies that can manage cognitive load in rich media learning contexts. The CTML builds upon three cognitive science principles: (1) dual-channel assumption (AKA, dual coding theory)–the human information processing system includes dual channels for verbal (i.e., auditory) and non-verbal (i.e., visual/pictorial) processing (Clark & Paivio, 1991); (2) limited-capacity assumption–each channel has a limited capacity for processing (Sweller, 1988); and (3) active processing assumption–active learning entails carrying out a coordinated set of cognitive processes during learning (Mayer, 2014). Based on these assumptions, CTML focuses on cognitive processes in multimedia learning, such as selecting, organizing, and integrating visual and auditory information for coherent knowledge representation (Mayer, 2014).

Data curation intersects multidisciplinary disciplines by nature where diverse communication modes can be integrated. Nevertheless, visual mode (graphs, tables, colors, and fonts) remains dominant in data curation practices among librarians and archive professionals due to its cognitive benefits for human learning and information processing as described in the CTML (Mayer, 2014). However, from this project, we argue that this widely accepted CTML based on dual (pictorial and auditory) coding theory needs to be redesigned carefully for blind and low-vision patrons and others having print disabilities, who would otherwise have to rely heavily on their verbal channel alone. For instance, we can compensate for visual information with other non-verbal modalities, such as non-speech sound effects and tactile shape representation in conjunction with verbal text mode.

2.2 Research Questions

Drawing upon the multimodal theories, this project aims to explore, suggest, and implement new ways of curating data that is accessible to people with varying degrees of sensory abilities. This three-year Early Career Research Development project will particularly focus on blind and low-vision population, considering both timeline feasibility and their

immediate needs for accessible solutions in visually-dominant data curation. To achieve this goal, this project will address the following research questions (RQs).

- Overarching RQ: How can we better support data curators to easily improve the accessibility of curated data (e.g., data charting and visualization) for blind patrons?
- RQ1: What challenges do data curators have when trying to create an accessible data visualization?
- RQ2: What challenges do blind patrons have when trying to interact with visually curated data and its information?
- RQ3: What design considerations contribute to inclusive and multi-modal data representations going beyond a single visual modality?
- RQ4: In what ways can accessibility be integrated into existing computational tools and reproducible frameworks as part of essential components for a sustainable solution?
- RQ5: How well does the suggested multimodal data representation interface address the accessibility challenges identified by data curators and blind patrons?
- RQ6: To what extent does the suggested multimodal data representation interface enhance accessibility and efficiency for both data curators and blind patrons?

2.3 Research Procedures and Methods

The project consists of three phases (each phase per year): Phase 1: needs Assessment and Solution Co-Design/Development; Phase 2: user testing and iterative refinement; Phase 3: public engagement and dissemination, which will be further described below.

2.3.1 Phase 1: Needs assessment and solution co-design/development (August 2023 - July 2024)

In Year 1, the PI and RA will conduct a needs assessment survey and interview to address RQ1 and RQ2 in collaboration with the two community partners (i.e., DCN and NFB). With the identified needs and challenges, the PI, the NCSA software directorate team, and the open-source partners (i.e., Posit PBC; C2M) will co-design and develop accessible solutions with a focus group of data curators and blind patrons to tackle the RQ3 and RQ4. The following describes detailed methods of data collection and analysis.

First, for the needs assessment, our team will obtain an IRB in month 1-2 and carry out an observational survey to investigate the current needs and challenges among data curators and blind patrons. Two end-user groups (i.e., data curators and blind patrons) will be the target subjects and they will receive a different set of questions addressing RQ1 and RQ2 respectively. For instance, data curators will be asked to (1) share their prior understanding and experience with serving blind patrons in their data curation work environments, (2) identify technical and training challenges addressing accessibility in their work, and (3) specify needs in regards to creating accessible data curation. Similarly, blind people will be asked to (1) share their prior experience with visually-curated data in their work and educational environments, (2) identify sources of challenges when interacting with visualized data curation, and (3) specify the types of data and curation services they wish to better access for their needs. In light of the exploratory nature of this study, convenient and snowball sampling strategies will be used and the survey link will be distributed through the two community partners (DCN for data curators; NFB for blind patrons). However, we will attempt to ensure diverse voices of each community are heard and included in our survey by reaching out to DCN's Racial Justice Special Interest Group and NFB's Diversity, Equity, and Inclusion Committee. A total of 200 responses (100 per community) will be collected and analyzed to reveal the current trends and needs, and suggest the design direction for our team's solution development.

Second, our team will recruit 2-4 experts from DCN and NFB, who will be involved in our co-design process to develop solutions addressing RQ3 and RQ4. Co-design is a collaborative process where both designers or researchers and non-designers work together to create a system (Sanders & Stappers, 2008). Co-design is a powerful form of human-centered design that incorporates the perspectives and preferences of target users into a novel tool, and recognized as one of the most effective approaches in ensuring that the needs of users are at the forefront of the design process

(Ladner, 2015). Although the number of co-designers may be subject to change that ranges from 2 to 4, we will follow the guidance of the local standards in the human-computer interaction (HCI) community (Caine, 2016). To minimize co-designers' burden, our team will host an hour co-design sessions before and after each major prototype iteration which is expected to happen bimonthly in Year 1. Additionally, it is worthy noting that the PI of this project is blind and an everyday assistive technology user, such as screen readers and refreshable braille displays. Moreover, we have Sandi Caldrone as our data curation consultant, who is experienced in curating data professionally. Both individuals are currently based at the University of Illinois at Urbana-Champaign. Their lived experiences, along with the co-designers' feedback, will drive the project direction towards a more feasible solution development.

Through iterative prototype development, our development team–consisting of the NCSA software directorate in close partnership with the Posit PBC and the C2M open-source team–aims to develop accessible and reproducible software that addresses RQ3 and RQ4 simultaneously. Considering the project timeline and broad impact, we will develop a cross-language data science package across R and Python that can translate each of the most widely used visualization objects, such as R ggplot (Wickham, 2010); Python matplotlib (Hunter, 2007), into accessible multi-modal (e.g., audible; touchable; readable) representations. The additional reason for choosing these two particular data science languages and visualization packages is due to the promising benefits of seamless integration with their literate programming and reproducible frameworks, such as Python Jupyter Notebook (Kluyver et al., 2016) and R Markdown (Allaire et al., 2023). Since each visualization package has their own graphic data structure, our development team will first design the abstract syntax tree (AST) 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 application programming interface (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 and his prior work. For example, the PI's research team have prototyped a front-end interface, called MAIDR (Multimodal Access and Interactive Data Representation; https://bit.ly/3YI5n8c) system and have been conducting pilot user studies. Recently, the PI and his MAIDR system were featured in Nature as a promising tool that enables visually impaired scientists to interact with scientific data, highlighting potential future opportunities (Katsnelson, 2023). 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, the PI 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; see the letter of commitment from Posit PBC for this evidence). 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). At the end of the Phase 1, the PI, RA and development team will conduct iterative and robust software tests (i.e., unit testing, integration testing, system testing, and accessibility testing) against the prototype before carrying out user studies in Phase 2.

2.3.2 Phase 2: User testing and iterative refinement (August 2024-July 2025)

In Year 2, with approved IRB study protocols, the developed toolkits will be evaluated and refined through iterative user study feedback from two end-user groups (i.e., professional data curators from DCN; blind and low-vision patrons from NFB) to address RQ5 and RQ6. While the Phase 1 will be focused on the co-design and development of accessible solutions reflecting the needs of the two user groups, the Phase 2 will be concerned with confirming whether the proposed solutions meet their needs and it is usable by them. For instance, 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 and reach the optimal granularity.

Using snowballing and convenient sampling, our team will recruit up to a total of 100 participants (50 from each community) to iteratively refine our tools' usability and accessibility. We will remotely perform three-round user studies with the balanced pairs of data curators and blind patrons at different times (i.e., 30-40-30 = n100 = 50 pairs) from which participants' feedback can be reflected towards the next refined iterations (see schedule of completion). For data collection and analysis, think-aloud protocols, semi-structured interviews, system usability scale (Brooke, 2013), and user burden scale (Suh et al., 2016)will be used. The PI and RA, the project development team and open-source partners will keep addressing identified issues and edge cases by bug fixes and patches until our system reaches a stable life cycle within the time frame. At the end of Phase 2, our team will also conduct a user acceptance test (Pandit & Tahiliani, 2015)with the co-designers who were recruited during Phase 1 to ensure that their recommendations and requirements have been met at an acceptable standard.

2.3.3 Phase 3: Public engagement and dissemination (August 2025-July 2026)

In Year 3, project outputs will be widely disseminated to the general public, , while the PI will strive to create a path for sustainable impact that extends beyond the scope of this project. The software packages developed, such as the JavaScript core engine, R binder, and Python binder, will be publicly released on GitHub as open-source projects under the GPL3 license. Detailed user manuals will also be provided, along with online sample galleries that will be made available under the CC BY-SA 4.0 license. To ensure easy installation and integration for the general public interested in using our tools for their projects, the PI and RA will release each software tool to widely adopted standard package management sites. For example, the visual-agnostic and multimodal JavaScript/TypeScript core engine library will be released on Node Package Manager (NPM), the R binder package ggplot2 on Comprehensive R Archive Network (CRAN), and the Python binder matplotlib on the Python Package Index (PyPI). More details can be found in the Digital Products Plan.

Online training,workshops, and webinars will be provided through data professional networks, including the Posit (FKA RStudio) annual conference, DCN events, and the Champaign-Urbana Data Science User Group. The PI and RA will spearhead the development and dissemination of open-source teaching modules on accessible data curation, which would be readily integratable into educational curricula of other faculty, librarians and data curators alike.

Furthermore, the PI and RA will make a dedicated effort to write and submit project results, which interweave with RQ1-6, to various national and international information conferences. These notable events include iConference, ACM ASSETS and CHI, CSUN Assistive Technology Conference, and the NFB Annual Meeting. By presenting at these leading conferences, our team will share our research and findings with a broad audience of scholars, researchers, industry leaders, and other professionals in the fields of library and information sciences, accessibility, and disability studies. After completing the final Phase of the project, the PI, RA, consultants, and advisory board will conduct a comprehensive assessment to determine the efficiency and impact of our dissemination strategies.

2.4 Project Personnel

The project team consists of the PI (Y1-3), RA (Y-3), developer (Y-2), and three technical partners (NCSA; Posit PBC; C2M) (Y-3), two community partners (DCN; NFB) for Y1-3, and 5-10 advisory board members (Y1-3).

The PI, Dr. JooYoung Seo, is an assistant professor in the School of Information Sciences (iSchool) and faculty affiliate at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign. Dr. Seo is also an internationally certified accessibility professional, RStudio-certified data science instructor, and one of the few blind scientists making accessible solutions as featured in the recent Nature(Katsnelson,

2023). His teaching and research involve accessible computing and inclusive data science for people with and without dis/abilities. Based on his prolific accessibility experience in research, teaching, and engineering, the PI will orchestrate the entire project directions to ensure every stage is ethical, accessible, and inclusive, following the proposed work plans. The PI will closely work with and advise an iSchool RA (Y1-3) for collecting and analyzing research data, and manage software development processes in collaboration with the NCSA Software Directorate (Y1-2).

The National Center for Supercomputing Applications (NCSA) is a pioneering academic institution located at the University of Illinois at Urbana-Champaign. It has been recognized as one of the most prominent research and development institutes where academia, industry, and government collaborate to solve grand challenges in diverse fields including computing, data, visualization, and cybersecurity. NCSA plays a vital role in advancing scientific discovery and breakthrough innovation through its cutting-edge computational resources and world-class expertise. As a technical partner, NCSA will provide the PI and his research team with necessary technical support, resources, and leadership to ensure successful project implementation.For example, Dr. Kenton McHenry (Associate Director for Software) and Dr. Jong Sung Lee (Associate Director) at NCSA will assist the PI in hiring and collaborating with a suitable developer for this project (job description is included with key project staff resumes).

Posit PBC (FKA RStudio) PBC is a technical partner that greatly contributes to the success of this grant. They have worked extensively on providing researchers and data professionals with an accessible, open-source framework for reproducible research (e.g., R Markdown; bookdown; knitr; Quarto), which aligns perfectly with the goals of this grant. In addition, Posit has also developed various open-source data science and visualization packages, including ggplot2 led by Dr. Hadley Wickham. Their support and expertise will enable us to achieve our objectives and drive forward progress towards more accessible and reproducible data curation practices. Dr. Tracy Teal (Open Source Program Director) and Dr. Carlos Eduardo Scheidegger (Software Engineer) will provide the PI's project team with technical consultation, resources, and outreach opportunities to a broader data science community.

Chart2Music (C2M) is an open-source team that creates an accessible data sonification solution for visually impaired individuals. The proposed project aligns well with their visual-agnostic approach, which translates data into sound. Dr. Sean Mealin, and Julianna Langston from C2M have agreed to collaborate with the PI and the research team to create accessible synergies towards multimodal data representation, which can be integrated into reproducible frameworks. They will provide technical consultation on visually-agnostic accessibility API design and data sonification expertise.

In addition to our technical partners, our team has established relationships with the Data Curator Networks (DCN) and National Federation of the Blind (NFB), allowing us to actively engage with our target beneficiaries throughout the project period (as detailed in the attached letters of support). As a formal DCN contact, Research Data Librarian and Assistant Professor at UIUC, Sandi Caldrone, will assist the PI in identifying five data curator advisory board members to provide constructive feedback on the project outcomes in Year 2. In addition, she will help disseminate the project outcomes to a wider data curator audience by co-hosting an online DCN workshop. Furthermore, the NFB will work closely with the PI during the recruitment of blind patrons through its various divisions and mailing listservs. The PI will participate in the NFB's Annual Event to present his work outcomes with other blind individuals who could potentially benefit from the developed solution.

Finally, Dr. Bongshin Lee, Senior Principal Researcher at Microsoft Research, will provide her prolific expertise and experience in data visualization to the project team (Y1-3) by serving as an advisory board member. Her constructive feedback will be valuable, and it will help in reflecting on the design and development process of the project.

3. Diversity Plan

This project actively addresses diversity, equity, inclusion, and accessibility (DEIA) 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. As a blind individual, Asian immigrant, and first-generation college graduate, the PI has extensive knowledge and related experiences that intersect DEIA and strives to promote its values through his research.

Throughout the project period and beyond, the PI's Accessible Computing Lab and the software development partner (i.e., NCSA) will be committed to following the diversity mission of the University of Illinois where everyone is welcomed, celebrated and respected regardless of their race, color, religion, sex, national origin, disability, sexual orientation, and gender identity. All of the technical partners (Posit PBC and C2M) prioritize DEIA in their open-source products that will also uphold our project mission.

Furthermore, we will identify and recruit participants with diverse backgrounds in collaboration with the two community partners (i.e., DCN; NFB). For example, DCN has Racial Justice Special Interest Group of data curators to create a more diverse, equitable, accessible, and inclusive environment. We will reach out to racially diverse data curators who can contribute to co-designing and testing accessible curation systems (Y1-2). We will also work with blind individuals with diverse backgrounds (Y1-2) because blindness impacts all races, ethnicities, and intersectionalities. The NFB has 25 divisions, 24 committees, and 9 groups that value diversity at the heart, such as National Association of Black Leaders, Committee on Diversity, Equity, and Inclusion, and NFB LGBT Group. Through close partnership with the NFB and their divisions, our team will ensure diverse voices of blind participants can be heard to create a more accessible and inclusive data curation system.

4. Project Results

The objective of this project is to develop an inclusive visualization tool and tutorials that enables data curators to provide crucial information to blind and low-vision people. Often, essential information is communicated through visual aids, and data curators lack tools and training to deliver visual information accessible. By addressing this gap, this project intends to produce several meaningful and sustainable results.

4.1 Development of Open-Source Tools and Tutorials

As a concrete outcome, this project results will lead to the development of multiple open-source data science packages that data curators and archive professionals can seamlessly integrate into their familiar data charting workflows and reproducible frameworks (e.g., Jupyter Notebooks; R Markdown; Quarto). Our visual-agnostic and multimodal API, R and Python packages will permit general data professionals with little experience in accessibility to easily turn their visualization into accessible data representations serving a wide range of patrons with print disabilities, including blind and low-vision people. This project will also produce user-friendly tutorials, online galleries, and community-based discussion forums on public GitHub repositories and pages where librarians, archive professionals, and data curators can learn from each other, share their tips and projects with the public, and improve their knowledge in accessible data curation. Furthermore, through collaboration with open-source community partners (e.g., Posit PBC; C2M team) as well as Data Curator network, the PI will develop and distribute open-source teaching modules on accessible data curation that other faculty, librarians, and data curators can easily integrate into their training and curriculum.

4.2 Positive Social Benefits to the Library and Information Sciences Field

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). It has the potential to change perceptions among data curators and librarians about accessibility, bringing up issues related to equality and justice that are central to the modern form of librarianship. Hence, the development of an accessible and open-source tool brings the potential benefits of a positive socio-technical impact towards inclusive librarianship and people with diverse needs. We will provide training sessions, webinars, and online tutorials for librarian and data curator participants to learn techniques and practices to produce accessible visualizations. The PI and his team will be actively presenting the project outcomes and results in national and international conferences (e.g., iConference; ACM CHI; ACM ASSETS; CSUN), data science meetups (e.g., posit::conf), and data curator summit (e.g., the DCN Next).

4.3 Sustainable Community Support

To ensure the project's benefits continue beyond its scheduled conclusion, we have put in place a few strategies. We will make the project codebase available under a GPL3 license and the tutorial galleries under a CC BY-SA 4.0 license as open source on public GitHub repositories. This step will enable users worldwide to use and improve the project, thus achieving continued growth and development. Additionally, our plans include integrating these tools into existing data curation workflows within libraries. This approach will save time, reduce resources needed for training and staffing while facilitating wide adoption among librarians and data professionals. As a result, these strategies will lead to sustainable and long-lasting community-based source code and tutorials, serving as an effective resource for the public. These long-term strategies will ensure that our project remains useful, accessible, and beneficial to the data curator community and users even after the completion of this project.

Schedule of Completion - page 1 of 3

Phase 1: Needs Assessment and Solution Co-Design/Development (Aug 2023-July 2024)

		Year One: Aug 2023 - July 2024											
		AUG	SEPT	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL
Obtain IRB for community needs assessment studies (e.g., surveys and focus-group interviews).	PI, RA												
Work with community partner organizations (e.g., DCN and NFB) to conduct a pre-survey for needs assessments on accessible visualization tools on a large-scale level.	PI, RA												
Recruit up to 4 experts (1-2 from DCN; 1-2 from NFB) who can co-design accessible solutions.	PI, RA, Community Partners												
Initiate development team setup, including a technical review on related work and engineer training.	PI, RA, NCSA Software Directorate												
Analyze data and identify needs, gaps, and opportunities for solution development.	PI, RA												
Co-design and co-develop solutions with 2-4 community partner experts and PI's development team with regular meetings.	PI, RA, Community Partners												
Develop a prototype solution that addresses identified needs.	PI, RA, Developer												
Hold a quarterly meeting to gather input from technical consultants and advisory board members for development.	PI, RA, Technical Consultants, Advisory Bd												
Conduct unit testing, integration testing, system testing, and accessibility testing against the prototype.	PI, RA, Developer												

***Note**: This is a tentative schedule. It is possible for some tasks to take a longer or shorter time to be finished. Testing of the software will be conducted along with the software development.*

Technical Consultants: Chart2Music open-source team; Posit engineers & research scientsts
Community Partners: Data Curation Network; National Federation for the Blind

Schedule of Completion - page 2 of 3

Phase 2: User Testing and Iterative Refinement (Aug 2024-Jul 2025)

		Year Two: Aug 2024 - July 2025											
		AUG	SEPT	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL
Recruit participants and develop user testing procedures.	PI, RA, Developer												
Initiate advisory board meeting.	PI, RA, Developer, Advisory Board												
Conduct first-batch user testing and collect feedback data from up to 30 participants (15 from data curators; 15 from blind individuals).	PI, RA, Developer												
Hold a quarterly meeting to gather input from technical consultants and advisory board members for development.	PI, RA, Developer, Tech Consultants, Advisory Bd												
Analyze user feedback and refine the solution based on findings.	PI, RA, Developer												
Conduct second-batch user testing and collect feedback data from up to 40 participants (20 from data curators; 20 from blind individuals).	PI, RA, Developer												
Hold advisory board meeting.	PI, RA, Developer, Advisory Board												
Conduct additional rounds of testing (15 from data curators; 15 from blind individuals) and refine as needed.	PI, RA, Developer												
Hold advisory board meeting.	PI, RA, Developer, Advisory Board												

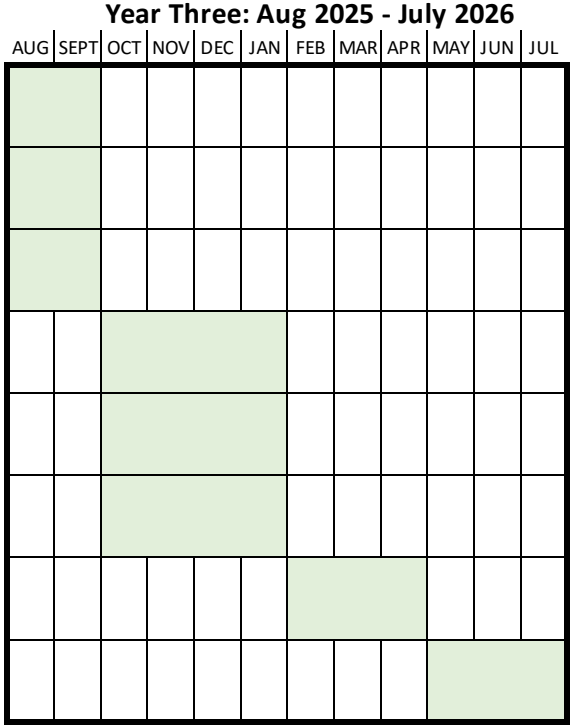
***Note**: This is a tentative schedule. It is possible for some tasks to take a longer or shorter time to be finished. Testing of the software will be conducted along with the software development.*

Technical Consultants: Chart2Music open-source team; Posit engineers & research scientsts
Community Partners: Data Curation Network; National Federation for the Blind

Schedule of Completion - page 3 of 3

Phase 3: Public Engagement and Dissemination (Aug 2025-Jul 2026)

Hold advisory board meeting.	PI, RA, Advisory Board
Develop a dissemination plan and identify target audiences.	PI, RA, Technical Consultants, Advisory Bd
Incorporate research into teaching/mentoring.	PI
Disseminate project tools through open-source platforms, including GitHub, NPM, CRAN, and PyPI.	PI, RA
Create publicly Accessible tutorials and galleries.	PI, RA
Write papers and give conference presentations.	PI, RA
Organize workshops, seminars, or webinars to engage with the public and showcase the solution.	PI, RA, Community Partners
Evaluate the impact and effectiveness of the dissemination efforts.	PI, RA, Technical Consultants, Advisory Bd



***Note**: This is a tentative schedule. It is possible for some tasks to take a longer or shorter time to be finished. Testing of the software will be conducted along with the software development.*

Technical Consultants: Chart2Music open-source team; Posit engineers & research scientsts
Community Partners: Data Curation Network; National Federation for the Blind

Digital Products Plan

Type

The research project aims to produce open-source packages in the following types and formats:

- Core engines: JavaScript/TypeScript libraries that can translate visual-agnostic JSON API into interactive non-speech sound, braille patterns, and text descriptions.
- R binder: an R package that can convert visual metadata from ggplot objects into visual-agnostic JSON API structure.
- Python binder: a Python package that can convert visual metadata from matplotlib objects into visual-agnostic JSON API structure.
- Tutorials and user galleries: HTML-based user manuals and examples that include detailed descriptions of each library and their respective functions, classes, and methods.

The type of research data may include interview transcripts (*.vtt; *.docx; *.txt; *.csv), audio and video recordings (*.mp4), survey feedback (*.csv; *.xlsx), system prototype design (*.png; *.svg; *.pdf), the project reports and publications (*.docx; *.gdoc; *.tex; *.pdf; *.html).

Availability

The project's open-source packages and materials will be publicly available through the following repositories

- JavaScript/TypeScript Core engines will be hosted on GitHub as a separate upstream repository and released in Node Package Manager (NPM).
- R binder will be hosted on GitHub as a separate repository and released in the Comprehensive R Archive Network (CRAN).
- Python binder will be hosted on GitHub as a separate repository and released in the Python Package Index (PyPI).
- Tutorials and user galleries will be hosted on GitHub public pages as well as Quarto Pub, which will be connected to an easy-to-remember custom domain name.

Data Management Plan specifies that depending on the sensitivity of the research data, different access and sharing policies are provided. For example, the human subject data will be only accessible to the research team during the project period. The project website and application servers will be hosted by the School of Information Sciences at the University of Illinois at Urbana-Champaign.

All the digital outputs will meet the Web Content Accessibility Guidelines (WCAG) version 2.1 AA compliance level for its universal access for people with and without dis/abilities.

Access

The PI and authors of the project's open-source tool will hold the original copyright. All the open-source tools that this project produces will be released under The GPL3 (GNU General Public License version 3) license to grant users the right to use, study, modify, and distribute software under certain conditions and to be made freely available to the public. Package tutorials and user galleries will be released under the Creative Commons Attribution-ShareAlike 4.0 International Public License (CC BY-SA 4.0) to grant permission to others to use, distribute, and build upon our work, subject to certain conditions.

Research papers published from this project will be posted to IDEALS https://www.ideals.illinois.edu/, the University of Illinois open-access repository. Use governed by the CC BY-SA license and will not have any access restrictions. Research interview data we generate will be governed by Institutional Research Board approval, gathered under a formal consent process, and stored under IRB-approved UIUC Box digital storage. Research interviews will be confidential

materials used internally for the development of accessible visualization tools, and will not be released beyond the project team.

Sustainability

Our team will oversee the entire open-source development process, including version control, issue tracking, testing, and release management. We will use GitHub public repositories, which allow anyone within or outside our project team to contribute towards developing the best possible solution during and beyond the project period. This means that people can suggest new ideas, report bugs, and create pull requests depending on their expertise in programming languages and software engineering tools.

Data Management Plan

Types of Data

The primary data generated in this project will consist of multiple datasets obtained through interviews, surveys, and experiments outlined in the project description.

Types of Software

The following proprietary applications are provided by the University of Illinois at Urbana-Champaign with educational license and will be used to collect, manage, and organize the data:

- Zoom conference for remote interviews, user studies, and hosting webinars.
- Qualtrics for collecting survey data.
- Box Drive for securely saving human subject data.
- Microsoft Office 365 (e.g., Word; Excel; PowerPoint; Outlook; Teams) for research communication.
- Google Suites (e.g., Google docs; sheets; forms) for team communication.
- Adobe Creative Cloud (e.g., Acrobat; Illustrator; Audition) for digital productions, including PDF, high-quality SVG figures, and audio files.
- GitHub Enterprise for source code management.

Additionally, we will employ the subsequent free open-source software:

- Visual Studio Code (VSCode) and its open-source extensions for general software development environments.
- RStudio IDE for a software development environment for R binders.
- Statistical- computing R environment for both qualitative and quantitative analyses, which can offer great accessibility between researchers with and without dis/abilities.
- Other open-source programming languages, such as Python, JavaScript, TypeScript, and node.js.

Notes: All the software mentioned above conform with the Americans with Disabilities Act (ADA) accessibility guidelines, and are either readily accessible or capable of being made accessible to individuals with disabilities.

Data and Metadata Format

In-process materials will be stored in Illinois Box for the use of the team. The interview will be recorded as *.mp4 and transcribed in *.vtt, *.txt, and *.csv formats. There will also be a note taker (*.docx). The dataset (i.e., transcriptions of the interviews) will not be publicly available.

Sensitive Information

The research activity requires IRB approval. It has not been approved. If the grant is received, the University of Illinois IRB approval will be secured. Full interview transcripts will not be shared in order to protect the privacy of our interviewees and to promote frank discussions with interviewees. For those participants that choose confidentiality on their consent form, any transcripts or recordings will be kept using a code system for identification. The written interviews/survey notes will be maintained in a secure file cabinet.

Policies for Access and Sharing

To ensure confidentiality, all data collected during the research will maintain anonymity via generated pseudo-random IDs. A code linking participant information to respective data will be kept in a secured cabinet at the University of Illinois at Urbana-Champaign. Only named investigators allowed under IRB policies will have access to shared data. The default handling of the data is through an encrypted server accessible exclusively via secure connections (e.g., https and ssh). Local access may be granted on a case-by-case basis, with investigators required to delete their copies post-analysis. Participant-approved explicit consent will govern data collection, which includes participant contact information.

Policies and Provisions for Reuse and Redistribution

The project aims to ensure public access by making the work products including technical reports and presentation materials available on the website. Besides, datasets will be made gradually available in an anonymous format without any personally identifiable information. Additionally, code that links data to specific participants and consent forms will be deleted after the specified retention period.

Plans for Archiving and Preservation of Access

To ensure data protection and integrity, we will perform daily backups of our data on separate servers at UIUC. Every quarter, we will also archive and store this data off-site in an encrypted format. Additionally, all research work products will be archived using standard UIUC practices and policies to maintain access. We anticipate that these data will continue to hold significant value for the research community even after the completion of the award-funded project. Hence, we plan to make them available on the project website to promote wider access and data reuse.

Software Sharing Plan

We prefer using open-source software as a means of spreading our system effectively to the research community. However, in particular situations, it may be more efficient to keep the intellectual property rights for parts or all of the software and make commercial use of resultant outcomes that can have wider impacts on society. Decisions related to the course of action taken will be based on the research outcome.

Data Management Plan Review

We will review the data management plan every quarter with the advisory board members. The PI will assess the status of the data collection, access, sharing, and archiving with the research assistants to ensure that the plan is executed as documented.

Organizational Profile

Founded in 1867, the University of Illinois at Urbana-Champaign (Illinois) is a non-profit, public land-grant university and is among the nation's most prominent research institutions. As a land grant institution, the University has a long record of commitment to public engagement and to the discovery and application of knowledge to improve and serve the greater society in which we live. As a world-class teaching and research institution, the UIUC campus provides vast resources to support the activities proposed.

It serves about 56,000 students, and employs approximately 1900 tenure-track faculty, 870 post-docs and specialized faculty, 2000 academic professionals and 1300 Civil Service staff. The students at Illinois represent all 50 states and 80 countries on average. The university offers 5,000+ courses in 150+ programs of study. An eminent STEM education and research hub, Illinois' ranking in the 2022-23 U.S. News' Best Colleges is 13 among public universities and 41 among national universities.

The School of Information Sciences

Founded in 1893, the School of Information Sciences (iSchool) is an American Library Association (ALA) accredited institution (renewed January 27, 2019) and a nationally recognized leader in the field, with a long history of supporting forward-thinking research and innovative academic programs that consistently rank as the best in the nation. The iSchool is currently home to the top graduate program in Library and Information Science, as well as #1 ranked specialty groups in Information Systems, Digital Librarianship, and Services for Youth, and three other programs ranked in the top ten. Much of this strength lies in the interdisciplinary expertise of the core faculty, who engage in work across more than forty broad-reaching research areas.

It is the School of Information Sciences' stated mission is *to lead the way in understanding the use of information in science, culture, society, commerce, and the diverse activities of our daily lives* (https://ischool.illinois.edu/our-school/strategic-plan). The iSchool is likewise dedicated *to shaping the future of information through* ***research, education, and engagement*** and *to creating sociotechnical solutions to real-world problems.*

To support that mission, the iSchool maintains an infrastructure of staff and technology dedicated to supporting our research agenda. The School likewise houses three preeminent research centers that serve as incubators for leading-edge research, including the Center for Informatics Research in Science and Scholarship, and publishes two distinguished periodicals, including the peer-reviewed journal Library Trends.

This interdisciplinary foundation, in turn, attracts a talented, diverse cohort of students at undergraduate, graduate, and doctoral levels. The iSchool currently offers 6 distinct degree programs, including the oldest LIS doctoral program in the country and 3 advanced certificate and licensure programs, all supported by an award-winning online education program. The iSchool is currently home to more than 1000 masters and doctoral students, and its undergraduate program includes nearly 450 students since its inception in Fall 2020.

As part of its engagement, the iSchool's researchers collaborate on interdisciplinary projects with scholars across our campus, including the University Library, and our local community, as well as institutions across the country and around the world. Local collaboration allows the iSchool to serve the Champaign-Urbana community and surrounding areas beyond the university itself. This service area includes the combined population of Champaign-Urbana, plus nearby small towns and rural communities (total population: approx. 230,000).

Organizational Placement & Governance: The School of Information Sciences is a unit of Academic Affairs, which is under the leadership of the Vice Chancellor and the Provost. Academic Affairs is one of the University of Illinois Urbana-Champaign's five top administrative offices.