Curating QDAS Research Objects for Exchange and Reuse

The Qualitative Data Repository at Syracuse University (QDR) seeks a 3-year National Leadership Grant for Libraries to study the sharing, reuse, and preservation of research objects produced by and dependent upon qualitative data analysis software (QDAS). The recent development of a data exchange standard for QDAS provides an exciting opportunity to study these problems, as well as to develop open source tools that enable the reliable preservation and reuse of QDAS research objects in data repositories. Dr. Sebastian Karcher and Dr. Colin Elman at QDR, Dr. Nicholas Weber from the Information School at the University of Washington, and Dr. Nathaniel Porter from Virginia Tech University Library, will address: 1) Which components of QDAS-based research are most readily shareable and re-usable and for what purposes? 2) How can effective guidance for researchers improve the shareability and reuse of qualitative data? 3) How can repositories take advantage of new developments in QDAS formats to facilitate the deposit and allow the exploration of such data? 4) How can shared QDAS data be used to enhance instruction in qualitative methods and QDAS tools?

National Need: QDAS empowers researchers to systematically analyze and code qualitative data. QDAS applications such as NVivo and ATLAS.ti are used widely across the health and social sciences. Correspondingly, academic libraries spend considerable resources in purchasing QDAS licenses and training users. QDAS research objects (the combination of data, annotations or codes, and resulting analytic outputs) are critical to qualitative scholarship, but they are rarely shared.This is despite the fact that there is now a widely held consensus that open science provides better outcomes for individual researchers, scientific communities, and for society at large. Similar benefits would follow from making available QDAS research objects, including codebooks and, where possible, fully coded projects. Sharing these data and materials would significantly improve the transparency of the large amount of qualitative research that depends upon QDAS, and at the same time produce invaluable resources for instruction.

The sharing of QDAS outputs has been blocked, partially, by the fact that most QDAS is proprietary, and each different QDAS application depends upon a unique data model (Corti & Gregory, 2011). Incompatibilities between these data models made it impossible to reuse or combine outputs between the different tools. However, recently the developers of leading QDAS applications agreed on REFI-QDA, an open exchange standard that allows the outputs of any QDAS application to be ingested in, and reused by, other applications. This new standard has the potential to catalyze an extraordinary increase in qualitative data sharing.

The REFI-QDA standard alone, however, will not bring about this result. Three other needs must also be met: First, QDAS user communities need **concrete guidance** for how researchers can best exchange, reuse, and receive credit for their QDAS research objects. Second, barriers to the exchange and reuse of QDAS research objects can also be greatly eased by the **development of open-source tools** that are integrated with existing scholarly research workflows, and that include managing, depositing, and exploring such data within trusted storage environments. Such tools would allow for easy exclusion of specific files or codes from a package prior to deposit and would allow online exploration of QDAS data at the repository. Third, we need to encourage uptake by **developing training materials** that help librarians to employ real world examples of QDAS research objects as they teach students how to use QDAS applications.

As the foremost social science repository in the United States with a dedicated focus on qualitative data, QDR is uniquely positioned to build on the unprecedented opportunity offered by REFI-QDA, and to address the national need for sharing QDAS outputs. QDR is expert in the use of QDAS, and in sharing QDAS outputs. Moreover, QDR is the only US-based repository advising REFI on configuring the standard to maximize its potential to archive QDAS projects. In addition, as a leading participant in dialogues about openness, QDR has a comprehensive understanding of the ethical and practical challenges of qualitative data sharing.

Project Design: Work under this grant will take advantage of significant previous *exploratory* work by QDR on the topic, including a workshop with practitioners and software developers (Karcher & Pagé, 2017) and several existing data projects based on QDAS materials. The bulk of the grant work will consist in *piloting* approaches to better sharing QDAS materials and then working towards *scaling* their availability and use. The **first stage** in the project will complement prior work with a systematic survey of researchers on their current practices and views on sharing QDAS data, building on and extending existing surveys (Xiao et al., 2014). The survey will provide a broad and rigorous overview of QDAS users, as well as identify potential depositors for collaboration.

With the help of an external selection committee with representation from diverse backgrounds, we will then invite a group of researchers to deposit their QDAS data, and we will support their additional work required with a stipend. These piloteers will work closely with QDR curators and keep a structured log of their activities.

The **second phase** of the grant will build on the survey and deposited data to conduct four related activities:

1) We know from previous work on QDAS archiving that a tool will be necessary for existing data repositories to ingest, publish, and make QDAS research objects meaningfully accessible. We will develop a REFI-QDA repository tool that streamlines depositing QDAS data into any Dataverse repository and allows for online exploration of shared data within a secure environment that can protect sensitive research products.
2) We will analyze the deposit and curation of the pilot projects, and develop guidance in the form of manuals and FAQs. QDR will coordinate the joint efforts of researchers, archivists, and instructors as they develop guidance materials, so that QDAS projects can be safely and effectively shared.
3) We will use the deposited data and guidance materials for instruction in different formats and venues (including both one-off workshops, classroom and graduate seminars) at the three partner institutions. Based on student and instructor feedback, we will further evaluate the characteristics of a maximally re-usable data project, and update the guidance materials accordingly.
4) The initial work teaching with deposited data will also lead to the development of a set of pedagogical instructions based on QDAS data, including lesson plans.

Diversity, Equity, and Inclusion: The project seeks to support diversity, equity and inclusion by integrating and supporting traditionally underrepresented populations in three groups: piloteers, data subjects, and data users. A minimum of one third of pilot projects will be by researchers from traditionally underrepresented groups, and we will work directly with researchers to ensure participation provides not just financial support but substantive benefit in furthering their research and visibility. Likewise, projects whose subjects reflect diverse populations will be prioritized for selection, with instructional material developed in collaboration with researchers. Finally, we will actively seek out and collaborate with instructors at under-resourced universities in developing and promoting instructional materials using shared QDAS data.

National Impact: The **final year** of the grant will shift to fostering wide adoption of the guidelines, technical tools, and instructional resources. The wider availability of shared QDAS projects will present a significant step towards more transparent qualitative research, allowing for qualitative data re-use, and greatly improving instruction in qualitative methods. The guidance for creating shareable QDAS projects will be published as a CC-BY-licensed whitepaper as well as a publication in a journal targeting QDAS users and presented at disciplinary conferences. We will also work with other repositories archiving qualitative data to implement versions of our guidelines. Our proposal to author a REFI-QDA curation primer has been accepted by the Data Curation Network (DCN) and we have ensured participation of a DCN mentor in the creation of the primer.

Software tools developed as part of the grant will be released under a free/libre and open source license. The tools will be integrated with QDR's Dataverse instance, we will support their adoption in other Dataverse repositories, facilitate adoption by other repositories. We will present the tools at Open Repositories and the Dataverse Community Meeting and provide detailed implementation documentation. QDR's technical team has a significant record of adding features to the Dataverse codebase, and we have received statements of collaboration from the Dataverse team at Harvard as well as the Global Dataverse Community Consortium to make QDAS-related enhancements available to all (currently 60) Dataverse installations worldwide.

Improving instruction in qualitative methods and tools is a key benefit of shared QDAS materials. To empower data and instructional librarians, who frequently teach and support QDAS, we will hold workshops on teaching QDAS with shared data at the IASSIST annual meeting and the RDAP summit and openly publish lesson plans.

Budget: We anticipate the total cost of the project will be $340,132, which includes no cost sharing. PI salary and benefits totals $38,334. Support for a graduate research assistant at Syracuse during year 1 and 2 of the grant will be $53,347. Costs for technical development are $90,000. Virginia Tech Libraries and the University of Washington, Seattle, will receive subawards totalling $104,451. Incentives for survey participation and stipends for 15 selected pilot QDAS projects will be $34,500. Conference travel for team members, particularly during year 3 or the grant is budgeted at $9,000, and an additional $10,500 will cover expenses such as catering and room costs for workshops during year 3.

Schedule of Completion

Year 1

	2021				2022							
Activities	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Stage 1												
Design and administer survey	■	■	■	■	■	■						
Analyze survey							■	■				
Recruit QDAS pilot deposits					■	■	■	■	■			
Begin deposit & curation of QDAS projects									■	■	■	■
Design initial archiving tool specifications		■	■	■								
Develop archiving tool baseline functionality					■	■	■	■	■	■	■	■

Year 2

	2022				2023							
Activities	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Stage 1 (cont'd)												
Finalize curation of QDAS projects	■	■	■									
Collect and file logbooks from depositors			■	■								
Stage 2												
QDAS tools: prototyping and deployment	■	■	■	■	■	■	■	■	■	■	■	
Analyze QDAS deposits and logs					■	■						
Write sharing CAQDAS (researchers)							■	■	■			
Develop REFI-QDA data curation primer							■	■	■	■	■	■
Teaching with deposited QDAS	■	■	■	■	■	■	■	■	■	■	■	■
Develop teaching materials										■	■	■

Year 3

	2023				2024							
Activities	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Stage 3												
QDAS tool implemented on QDR	■											
Refine QDAS tools at QDR; support other implementations		■	■	■	■	■	■	■	■	■	■	■
Publish and present on QDAS tool								■	■	■	■	
Promote QDAS Sharing guidelines	■	■	■	■	■	■	■	■	■	■	■	■
Present on sharing QDAS; write & publish whitepaper; write & submit journal article			■	■	■	■	■	■	■	■	■	
Publish data curation primer	■	■										
Promote and refine data curation primer			■	■	■	■	■	■	■	■	■	■
Final report and summary of grant outcomes												■

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

INSTRUCTIONS

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS
Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE
Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA
Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). Explain and justify any prohibitive terms or conditions of use or access, and detail how you will notify potential users about relevant terms and conditions.

All software produced under this grant (QDAS tools) will be released under an open source license. Software integrated with Dataverse will be released under an Apache License (version 2.0). Other software will be released under an MIT license, unless included open source components require the use of a different open source license (such as GPL). In all cases, the Qualitative Data Repository will be the formal holder of copyright for software.

All pedagogical materials, as well as technical documentation and primers will be made available under CC-BY license. The copyright remains with the materials' authors.

Subject to participant consent, all data generated through this grant (survey and pilot logs and reports) will be made available for research and teaching through QDR under the repository's standard terms which allow for free use for teaching and research.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

All digital products will be distributed under permissive, open licenses. The only product with significant restrictions are the de-identified human-participant data produced as part of the grant. These data will be freely accessible for research and teaching but will not allow for commercial use or redistribution, in line with QDR's standard terms and conditions to protect human participants.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

The data collected as part of the survey and, especially, from the QDAS piloteers may raise some, albeit limited, privacy concerns. Both activities will be conducted under IRB protocol (see below) and participants' data will only be shared beyond the project team where they explicitly consent to such data sharing.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

We will create two principal forms of digital resources: 1) teaching resources and lesson plans for teaching with QDAS data and 2) a data curation primer for REFI-QDA. We plan to create a minimum of four different lesson plans, initially composed in google docs and distributed in PDF and HTML format. In line with the Data Curation Network's standard practice the data curation primer will be written in markdown and made available via the DCN's github repository and can easily be converted into other formats (PDF, HTML, etc.).

A.2 List the equipment, software, and supplies that you will use to create the digital content, resources, or assets, or the name of the service provider that will perform the work.

Resources will be generated by the content team with standard tools: google docs, github, markdown, and potentially a static website generator such as jekyll to provide an easy, visually attractive online presence. Such a site would be generated by the project team using github pages and a freely available jekyll template.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

Resources will be text-based and use simple, standard formats: md, PDF (PDF/A compliant), and html.

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

All resources will be reviewed by multiple projects staff before publication. We will use github issues and products boards to track and monitor product development, including templated checklists to support quality control. We will also update all products based on post-publication feedback. Workflows and technical details of published products are purposefully simple and do not require additional quality control.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

Digital assets (other than data and software, specified below) will be deposited in Virginia Tech's VT Works (https://vtechworks.lib.vt.edu/). For any teaching resources that involve more complex content, such as the website listed under A.2, we will use robust technology (such as static webpages) to limit the need for maintenance. We will also make any source code available via github and archive the github repository using Zenodo on project completion to ensure long-term preservation of assets.

Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata or linked data. Specify which standards or data models you will use for the metadata structure (e.g., RDF, BIBFRAME, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

Assets created under this project are purposefully simple to preserve. They will be described using Dublin Core metadata to maximize findability.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

Metadata will be stored in VT Works and Zenodo alongside preserved assets.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Zenodo and VT Works expose Dublin Core natively and support standard harvesting protocols to support finding.

Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content, delivery enabled by IIIF specifications).

Digital content will be available openly to everyone websites (a dedicated site for pedagogical content and the DCN's site for curation primers) and preserved in digital repositories as described above.

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

Managing Qualitative Social Science Data: https://managing-qualitative-data.org/
QDR Guidance: https://qdr.syr.edu/guidance

SECTION III: SOFTWARE

General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

The QDAS archiving tools will read and understand REFI-QDA's XML format, unzip the contents of the .qdpx project files and decide how they are best stored on the repository, and build basic explore functionalities based on .qdpx files. Both functions (deposit and explore) will be integrated with the dataverse software platform. They are intended to be used by researchers to deposit QDAS projects in data repositories and to explore such projects online. The tools are intended to be installed by other data repositories.

A.2 List other existing software that wholly or partially performs the same or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

The dataverse software provides comparable archiving and explore functionality for quantitative data. Spreadsheet data is automatically normalized tab-delimited format on deposit and explore tools like Two Ravens (http://guides.dataverse.org/en/latest/user/data-exploration/tworavens.html) and Data Explorer (https://github.com/scholarsportal/Dataverse-Data-Explorer) allow for data analysis and exploration of such data.

No comparable facilities, either for archiving or for exploration, exist for QDAS data for any repository platform. The proposed tools help to fill this gap.

Technical Information

B.1 List the programming languages, platforms, frameworks, software, or other applications you will use to create your software and explain why you chose them.

The tools' backend will be coded in enterprise Java. This ensures maximal compatibility with the dataverse platform, as well as access to Java's unparalleled set of libraries for document manipulation such as pdfbox. Java is also the most widely used programming language for other data repository platforms such as Fedora and Dspace.

The frontend of the tools, i.e. both the deposit and the display interface, will use standard html and javascript/jquery.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

As described in more detail in the proposal, we will focus our efforts on Dataverse integration. Dataverse has existing options for external tools and viewers, facilitating such integration. We will aim to make software components modular to allow for easy integration into other tools.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

None.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

We will develop the QDAS tools publicly on QDR's github repository. During stage 2 of the grant, we will employ rapid prototyping and deployment to incorporate user feedback. All software will include both usage and installation instructions, either on github or on readthedocs. We will create these instructions concurrently with development and have dedicated time and resources during stage 3 of the grant to refine documentation and assist other instances in implementation.

QDR has a proven track record of creating open source tools for wider usage, such as the "Dataverse Previewers" https://github.com/GlobalDataverseCommunityConsortium/dataverse-previewers

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

Dataverse previewers: https://github.com/GlobalDataverseCommunityConsortium/dataverse-previewers
archivR (R package):
https://github.com/QualitativeDataRepository/archivr
Managing Qualitative Data (jekyll page for online course):
https://github.com/QualitativeDataRepository/ssrc-qdr-data-management
QDR's dataverse instance:
https://github.com/QualitativeDataRepository/dataverse

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

The code will be accessible through a public github repository, which will also include installation instructions

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

Qualitative Data Repository Github

URL:

https://github.com/QualitativeDataRepository

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

A.1 Identify the type(s) of data you plan to collect or generate, and the purpose or intended use(s) to which you expect them to be put. Describe the method(s) you will use, the proposed scope and scale, and the approximate dates or intervals at which you will collect or generate data.

The data will include survey data with about 500 observations, collected via online qualtrics survey and saved locally as .csv files as well as the logs and reports by pilot depositors, 15 each, initially collected as Word (.docx) or PDF files.
As part of the project, other researchers will also deposit 15 QDAS based qualitative data projects with QDR. These data themselves, however, will have been created outside (and likely before the start) of this grant.
All data will be collected during the first 18 month of the grant (see the schedule of completion for details.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

The survey and likely the collection of reports and logs do qualify as human participant research and will require IRB approval. Given the minimal risk to participants (who report on their every day activities) we expect the IRB to exempt our application for review. We will apply for IRB approval in June-July 2021.

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

We will collect identifying information (email, name, and position) as part of the survey as well as the pilot activities. We will remove identifying information from the survey before sharing.
The pilot reports and logs will be impossible to de-identify given the amount of contextual information included. We will seek participants' permission to share these with their name included.
None of the information included in survey or reports/logs is sensitive in nature: respondents report on their usage of digital tools and views on data sharing.

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

The survey data will be in standard tabular format and can be opened in any spreadsheet or data analysis tool.
The logs and reports will be provided as PDF/A.

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

The documentation will include consent agreements, detailed documentation about recruitment for both the survey and the pilots, the codebook for the survey, as well as instructions provided to piloteers for reports and logs.

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

All data will be shared through the Qualitative Data Repository (QDR) within 12 months after project completion or with the first formal publication based on project results, whichever comes earlier. QDR is a Core-Trust-Seal certified repository specializing on qualitative and multi-method data such as those produced under this grant. Access to all data in QDR is free for everyone.

A.7 Identify where you will deposit the data:

Name of repository:

Qualitative Data Repository

URL:

https://data.qdr.syr.edu

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

he DMP will be part of the project documentation and be converted into a google doc to make it a living document. Dr. Karcher will monitor adherence to the plan and it will be reviewed at a minimum every 6 months or when questions/concerns arise.