

Annotation for Transparent Inquiry (ATI): Demonstrating and Evaluating Open Annotation for Publications Based on Qualitative Research

Achieving openness in qualitative research is difficult because of the dense connections between data, analysis and text. Quantitative analysis precedes writing, and results enter articles in the form of summary tables and their discussion. Qualitative analysis, by contrast, often both precedes *and* occurs concurrently with writing, and data are woven throughout the text of a book or article. Open annotation – which allows for the generation, sharing, and discovery of digital annotations across the web – allows scholars to provide requisite information about how their data were generated and analyzed, empowers annotations to be treated as FAIR data, and retains and enhances the readability and impact of scholarly research.

The Qualitative Data Repository (QDR, www.qdr.org) and Hypothes.is are deploying an open annotation server to host data supplements. We have also begun to collaborate with prominent academic and scholarly publishers to identify pilot projects that will illustrate the power and promise of open annotation. Our overarching goal is to demonstrate and promulgate the utility of a technique for achieving transparency developed by QDR: “annotation for transparent inquiry” (ATI). This letter of interest requests support to broaden our outreach efforts by enlisting additional pilots, further publicizing the availability and virtues of the open annotation solution, evaluating the pilot projects, and identifying gaps and providing guidance for the next stages of development.

The problem we are trying to solve

Quantitative analysis uses numeric data arranged in a matrix and approached as an aggregate body of information. Analysis precedes writing and is reported in summary form in a book or article. Qualitative data generation and analysis, by contrast, are granular and emergent. **Granular** qualitative data are considered individually or in small groups; the content of each cited source (e.g., book, interview transcript, newspaper article, video clip, etc.) serves as a distinct input to the analysis. Moreover, the conclusions reached in qualitative inquiry are often **emergent** from and coterminous with the writing. Data, analysis, and conclusions are densely woven across the span of a book or article

The problem QDR and Hypothes.is are trying to solve is optimizing openness in qualitative research, by simultaneously satisfying several desiderata:

1. In order for the article to satisfy **openness criteria**, the solution has to provide the requisite information about how the data were produced and analyzed.
2. The additional material should be **FAIR data**—i.e. findable, accessible, interoperable, and reusable.
3. The solution has to **digitally enrich** individual articles, making the reading experience more compelling. For example, where possible, data sources should be immediately available, especially those that are intrinsically interesting and important.

Satisfying these desiderata requires the participation and coordination of different types of institutions. The qualities of granularity and emergence suggest that data must be deployed – and thus annotations integrated – across the span of an article. As such, the annotations must be linked to the article. The additional content annotations provide serves as a digital exoskeleton, dramatically strengthening the original article, enhancing its evidentiary base and increasing the amount and quality of information about the analysis. The digital content augments the main text and notes and may obviate the need for block quotations and discursive footnotes (thereby opening more space in the article for substantive content).

The FAIR requirement insists that qualitative data sources be findable, accessible, interoperable, and reusable. FAIR imposes responsibilities that go well beyond what most publishers are willing or able to deliver. This suggests the need for a third party to host annotations and underlying data sources in a sustainable venue where attention is paid to long term usability (e.g., providing updated versions of files when the software designed to access them changes). Repositories are not, however, in a position to publish the articles to which the annotations are linked.

Our central challenge, then, is to devise a solution that allows stakeholders to do what they do best – publishers to publish articles, repositories to host FAIR data, and technology firms to build annotation platforms – and that facilitates the tight integration of these activities, all of which are required for openness in qualitative research.

The methods that will be used

QDR and Hypothes.is are deploying an open annotation server to host data supplements – and ATI projects in particular – accompanying articles published in leading social science journals that publish on multiple platforms. In essence, the open annotation server acts as a coordinating device: it serves open annotations that are mapped to an article and linked to underlying data sources in QDR. Authors can use annotations to comment on and clarify individual claims and to link primary documents (interview transcripts, archival images, etc.) to those claims.

Open annotations solve the publisher-repository structural problem. Annotations can be presented on any web page from a third party annotation server. Displaying annotations requires client software, and the most widely used such software is provided by Hypothes.is. Annotations become visible on the click of a button – a Hypothes.is bookmarklet or a Hypothes.is Chrome add-on (which require no action or participation by journal publishers) – or an enabled button on a journal web page. When annotation is enabled, annotated passages in an article published on a journal web page are highlighted and annotations are visible on the right side of the web page. Anyone with access to the article can open the annotations. Moreover, because the annotations sit on an external server, publishers have no curation costs. Each stakeholder does what it does best, while gaining maximum benefit from the other's efforts. Publishers will continue to publish articles and books, and data repositories such as QDR will curate, hold, and distribute data in the form of open annotations.

Open annotation satisfies the three desiderata mentioned above. First, annotations provide extensive real estate for authors to provide information about how the data were generated and analyzed. Second, annotations are treated as data, and the collection of annotations associated with a particular article are viewable as a single data supplement and assigned a DOI. Individual annotations and underlying data sources will also have individual DOIs, allowing for specific, machine-readable citations to any annotation or data source. The annotations and the underlying data sources are searchable via the repository catalog. Finally, annotations are mapped to the text of an article, with underlying data sources immediately available.

The impact that will be achieved at the completion of the project

Upon project completion, open annotation will have seen significant use and study as a means to facilitate transparent qualitative research. If invited to submit a full proposal, we anticipate requesting \$200,000 to produce the following outputs and associated impacts:

1. In order to show the promise of open annotation, we will publish 20 pilot ATI projects accompanying 20 articles in various social science and related disciplines. We have already begun to collaborate with prominent academic and scholarly publishers to identify promising pilot projects. The pilot projects will popularize ATI as a technique for achieving openness, and encourage researchers to think of annotations as qualitative data. Both outcomes should lead to wider adoption of openness and data transparency in fields that have been reluctant to embrace these practices. In collaborating with journal publishers, we will gain insights, both practical and technological, on workflows for adding and promoting annotations-as-data and on best practices in curating ATI projects. These insights will be published both as a technical whitepaper and in a data science journal, so that any findings can be used and applied widely, including by other open science projects that build on open web annotations.
2. Hypothes.is will provide critical technical support in ingesting and serving the annotations. In addition, we will fund and provide access to an information science researcher to document and evaluate the process of data publication, and. The results of this assessment should help us to understand to what extent ATI is able to fulfill its promise to make qualitative research more transparent, and to provide specific recommendations for possible adjustments to the data format and workflows. Any outputs of this research will be published open access.
3. An increasing number of journals are requiring quantitative analyses to be verified before publication – either in a fully automated way, or by relying on third party institutes. While fully automated verification will likely remain impossible in the foreseeable future, the development of the pilot ATI projects will allow us to evaluate if, to what extent, and at what costs verification of qualitative analysis is possible. As verification policies become more common and codified (e.g., in the Center for Open Science’s [COS] Transparency and Openness Promotion [TOP] guidelines), our findings will be an invaluable guide to their adoption for qualitative research.
4. Lastly, as we assemble and treat ATI projects as standalone data, researchers will be able to explore secondary analysis of such data. Whereas secondary analysis is one of the major drivers in the push for data sharing in the quantitative tradition, to date, secondary analysis of qualitative data is exceedingly rare. A corpus of qualitative data, collected and displayed in the same format would provide unique opportunities for secondary analysis, e.g., of methodological practices and, as more ATI Ps becomes available, within given issue areas.

A statement describing why we are interested in “openness” and any previous work on this topic.

The co-PIs on the present proposal direct QDR, which is funded by the National Science Foundation, and Hypothes.is. QDR stores, preserves, and publishes digital data arising from qualitative and multi-method inquiry. Hypothes.is [short text on Hypothes.is to be added here.....].

Qualitative research makes important contributions in both the natural and social sciences. While its virtues are more commonly acknowledged in social science, qualitative research is (for example) also indispensable in the essential exploratory work that often precedes causal identification in medical research, e.g., via case reports. Qualitative data and analysis are vital to solving social, political, economic, environmental, and health problems important to our collective well-being and security. While also allowing data to be re-used for secondary analysis and teaching, *the core justification for openness is rigor*. Science’s claim to produce privileged knowledge is procedurally contingent. Transparency is a logical concomitant of rigor. *It is impossible to make claims for the value-added of science as a producer of privileged knowledge without simultaneously supporting openness.*

QDR and Hypothes.is take the view that the openness imperative applies across the natural and social sciences and extends to multiple techniques for generating data and multiple methods for analyzing them. As a consequence, we have been fully engaged in the recent reshaping of data and publishing infrastructure to provide mechanisms for making inquiry more transparent. We have also helped revise research practices and standards, to provide scholars with professional incentives to engage in the kinds of effort that help others understand the content and meaning of their conclusions. In addition to leading their own prominent organizations, project participants have published widely on transparency, openness, and the challenges of managing and sharing qualitative data; they also play important advisory roles in cognate organizations (e.g. COS, the Inter-university Consortium for Political and Social Research [ICPSR], and the Data Access and Research Transparency [DA-RT] initiative in political science).