

Annotation for transparent inquiry: Transparent data and analysis for qualitative research

Sebastian Karcher¹, Nicholas Weber²

Abstract

How can authors using many individual pieces of qualitative data throughout a publication make their research transparent? In this paper we introduce Annotation for Transparent Inquiry (ATI), an approach to enhance transparency in qualitative research. ATI allows authors to connect specific passages in their publication with an annotation. These annotations provide additional information relevant to the passage and, when possible, include a link to one or more data sources underlying a claim; data sources are housed in a repository. After describing ATI's conceptual and technological implementation, we report on its evaluation through a series of workshops conducted by the Qualitative Data Repository (QDR) and present initial results of the evaluation. The article ends with an outlook on next steps for the project.

Keywords

qualitative data; research transparency; annotation; digital technology

The Problem: Transparency for Qualitative Research

Research transparency is rapidly becoming the norm in empirical social science scholarship. Journals (e.g. Giofrè et al. 2017), funders (although see Couture et al. 2018), professional organizations (e.g. APA 2016; APSA 2012), and peers (e.g. Freese and King 2018) increasingly expect research to be transparent. In its ethics guidelines, the American Political Science Association (APSA) distinguishes between three components of transparent research (APSA 2012): production transparency, data access, and analytic transparency.

For quantitative empirical research, a basic template has become well established on how to provide both access to data and analytic transparency (King 1995). Authors enable others to reproduce their findings by sharing data along with the computer code used to produce the analysis presented. A similar template, however, does not exist for qualitative research. This article presents Annotation for Transparent Inquiry (ATI), a solution for making work based on qualitative and multi-method data more transparent. ATI allows scholars to annotate specific passages in a publication in order to provide extended commentary on arguments made in the text and, when possible, include a link to one or more data sources underlying a claim. Linked data sources are stored in a repository. ATI presents a unique mix of *methodological* and *technological* solutions to the problem of transparency for qualitative research.

In the following paper, we begin by describing ATI's origins and its conceptual basis. We then outline the technical implementation and the considerations guiding the choice of technology to implement ATI. In a third section we describe our process for evaluating ATI based on a set of commissioned pilot studies

and their review by subject-area experts. We present some initial results from this evaluation and conclude by outlining next steps for ATI.

Introducing Annotation for Transparent Inquiry

Calls for greater transparency in qualitative research are not novel. In a series of articles, political scientist Andrew Moravcsik (2010, 2014) advanced the idea of “active citation” in which “any empirical citation be hyperlinked to an annotated excerpt from the original source, which appears in a ‘transparency appendix’ at the end of the paper, article, or book chapter” (Moravcsik 2014, 50). Moravcsik’s idea was taken up by the newly founded Qualitative Data Repository (QDR), which commissioned eight authors to pilot a dedicated interface for displaying and viewing active citations (e.g. Crawford 2015; Saunders 2015). These pilot projects demonstrated the appeal of using active citation to improve qualitative transparency, but also highlighted limitations of the approach. Most importantly perhaps, the active citation viewer developed by QDR required us to obtain copyright permissions from the original publisher of the article in order to create and host a copy of the article separate from the authoritative published version. Moreover, active citation was poorly connected to existing standards and trends in scholarly digital publishing. It relied on custom software and relied heavily upon online appendices, at a time when those are falling out of favor, especially for research data (Kratz and Strasser 2014). By lacking a clear location for shared primary data (e.g., interview transcripts or archival scans), active citation also raised concerns about the digital preservation of such data.

Based on these concerns, QDR sought to develop a new approach to qualitative transparency that could make use of existing open-source technologies, draw upon emerging web-based standards for scholarly publishing, and create a more robust link between active-citations and their underlying data. In the following sections we describe the conceptual and technical design of this approach, which we have called Annotation for Transparent Inquiry (ATI).

Conceptual implementation

Using ATI, authors can annotate any passage in their work that they wish to add additional information to. Every annotation includes at least one (and possibly all) of the following elements (see <https://qdr.syr.edu/ati/ati-instructions>):

- A **source excerpt**: typically 100 to 150 words from a textual source (e.g., an excerpt from the transcription for handwritten material, audiovisual material, or material generated through interviews or focus groups);
- A **source excerpt translation**: if the excerpt is not in English, a translation and indication of its source;
- An **analytic note**: discussion that illustrates how the data were generated and/or analyzed and how they support the empirical claim or conclusion being annotated in the text;
- **Data source**: a link to the underlying data source that can be shared legally and ethically (currently QDR, but potentially any trusted digital archive);
- **Full citation**: any additional details to describe the source being excerpted or linked and its location.

The close link between text and data provided by annotation, both visual and in terms of the data structure, corresponds to the close interaction between text and data in most qualitative research. The

linked data sources of an ATI project are housed in a data repository that also provides the landing page for the data project as a whole as well as additional documentation (see Figure 1). Beyond standard metadata (keywords, dates, geographic tags, funding, etc.) this landing page also includes a data overview, in which authors explain both general background information about their data (collection strategy, provenance, etc.) and their “logic of annotation.” The logic of annotation provides key information to readers on what types of annotations to expect and where: Authors may, for example, focus on annotating controversial claims or claims that are of great importance to their argument. They may also choose to use annotations to provide additional context or “color” to particular passages -- which may be of particular interest for researchers working in more interpretivist traditions (e.g., ethnographers who seek to “thicken” a textual description, following Geertz 1973). While QDR suggests some options for this, the choice of logic is left to researchers to allow for the wide variety in methodological approaches within qualitative social science.

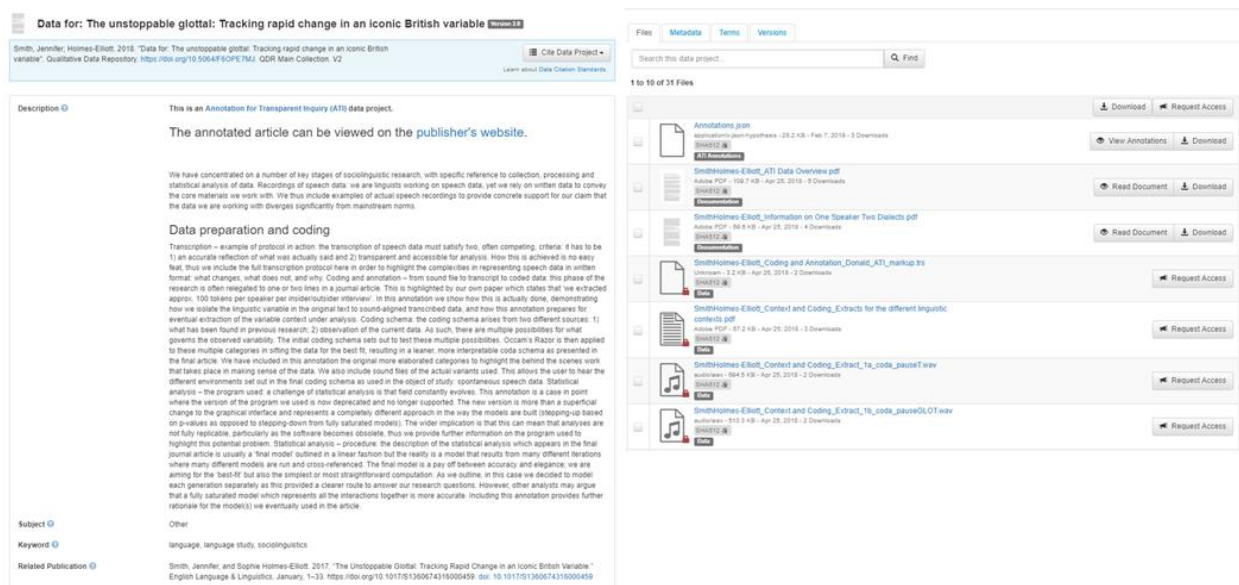


Figure 1: Data landing page for an ATI project (here: Smith and Holmes-Elliott 2018)

The initial annotation on the article links to the data landing page, and the data landing page likewise features a prominent link to the article with annotations. This bidirectional linking ensures that readers can easily find and navigate both components of the ATI data project. The data landing page is also where a Digital Object Identifier (DOI) for the data directs, allowing for separate citations to the data. A set of published pilot studies is available at <http://qdr.syr.edu/ati/ati-models>.

Technical Implementation

One of the major concerns in asking researchers to use new methods is the need to use and learn new tools, which are frequently a source of frustration. QDR therefore opted to *not* impose any given annotation technology on depositors. Instead, authors can generate ATI annotations using their own choice from familiar tools including comments or cross-references in Word or LibreOffice, comments in Acrobat, hyperrefs in LaTeX, etc. These formats are then converted into web annotations by QDR.³

Annotations are displayed alongside the article using a javascript-based open-source software tool called Hypothesis (<http://web.hypothes.is>). The annotations are accessible in a ‘public group’ (Walker-

Peddakotla 2018) that provides a layer visible to any reader worldwide but only editable by QDR personnel. Based on articles' DOI, annotations are visible across identical versions of the same article. Similarly, annotations are visible on both PDF and HTML versions of articles (see Udell 2017 for details and examples).⁴

Hypothesis provides an ideal technical platform for transparency: the software is open source under a simplified BSD license (Hypothesis 2019), the data and annotation model is based on a recently accepted W3C web standard (Sanderson, Ciccarese, and Young 2017), and Hypothesis itself is a non-profit whose goals align closely with those of open science advocates. Given the open nature of the tools and the annotation data, switching to a different annotation client or supporting multiple clients will remain feasible, reducing dramatically the chance of lock-in.

The Qualitative Data Repository has recently developed a tool integrated with the Dataverse repository software to ingest annotations via the Hypothesis API. This tool allows curators at QDR to enter the identifier for a set of annotations and then import a JSON file with the annotation content, text anchor, and other relevant information. QDR's implementation of this tool also transforms the JSON file into an HTML page. This functionality enables a viewer of the ATI project on QDR to see the full set of annotations that are related to a project without having to access the full article on a publisher's website. This feature also satisfies a use case where a QDR user is searching or browsing the repository and discovers a new ATI project. The user can then view all related data files and the annotations (as well as anchored texts) in one single, convenient location.

Evaluating ATI through Workshops

The Qualitative Data Repository evaluated both the conceptual and technical implementations of ATI through two workshops held in 2018. These workshops were sponsored, in part, by grants from the Robert Wood Johnson Foundation and the National Science Foundation.

Participants

In the first workshop, participants were recruited from two populations: A set of authors were invited to use ATI to help improve the analytic, data, or production transparency of a recently published journal article. These participants were identified by recommendation from editors of journals, due to their topical expertise in qualitative data analysis, or by identifying unique use cases where ATI might improve the access to underlying data. Once the authors had been recruited, they were asked to suggest a graduate student or early career researcher to act as a reviewer of their ATI project.

For the second workshop, authors developed an annotation scheme to improve the transparency of works in-progress instead of applying annotations to previously published articles. To recruit authors of in-progress research QDR created the "ATI Challenge," which acted as a call for proposals from early-career researchers. Applicants to the ATI Challenge were asked to describe how and why they planned to use annotations to improve the transparency of work that they were preparing for publication in a peer-reviewed journal. A selection committee of qualitative research experts identified the most innovative proposals. Similar to the first workshop, authors then suggested a topical expert that would act as a reviewer of their annotations (see Table 1 for a breakdown of participants across the two workshops).

Evaluation

In advance of both workshops, authors were asked to annotate their research articles with the goal of improving the narrative through analytic, data, or production transparency notes. This included depositing relevant data with QDR as described above. Reviewers were asked to first read an article and mark passages where they would *expect* to see annotations or underlying data. Reviewers then re-read the article with the authors' annotations to evaluate how annotations improved the transparency of the argument that was being advanced. Reviewers prepared a summary of their experience reading both the original and annotated version of an article, and provided written and oral feedback to the authors at the workshop.

Data Collected from Participants

QDR asked the authors to keep a written log explaining the procedure they followed for selecting and applying annotations to their articles. Authors also provided feedback on their experience as producers of ATI through both a pre-workshop survey and a set of in-person focus groups at each workshop. Reviewers were similarly asked to answer detailed questions on their experience as consumers of ATI, and similarly participated in focus groups at each workshop. Below, we present some preliminary observations based on this data.

	Author / Annotator	Reviewer	State of Manuscript
Workshop 1	Established Expert	Early Career	Published
Workshop 2	Early Career	Established Expert	In Preparation

Table 1: ATI Initiative Workshops

Reactions to Producing and Consuming ATI

Participants at both workshops expressed enthusiasm for access to tools and techniques that support a more transparent qualitative research paradigm.⁵ Authors reported that using ATI provided more space to explain and defend their use of a method or a particular data source, and an overall sense of liberation from the limited word-counts that journal publishers often impose. Authors also reported that they felt their overall argument was improved by having the ability to justify not just what was *included* in the narrative, but why certain methods were *not* used or why some data had *not* been included or analyzed. Reviewers reported an overall positive experience reading and consuming annotated articles. Many reviewers observed that the annotations provided answers to initial questions they had posed about why a research design or analytic method was chosen. However, with more insight into these choices the annotations opened up new questions and lines of critique from reviewers. In many ways this should be viewed as a positive outcome -- through improving production transparency and easing data access, the reviewers were able to focus more specifically on the arguments that were being advanced by an author's data analysis.

Participants also identified a number of challenges facing the broad uptake and use of ATI. These challenges can broadly be characterized as uncertainty in selecting what to annotate, incentives for producing ATI projects, and the clear identification or signalling of different types of annotations to consumers.

Annotation Selection

Authors who applied ATI to already completed research articles faced the challenge of having to revisit previous research notes, locate data sources, and defend causal claims that were developed months and in some cases years in the past. In short, the revisiting of completed research articles was instructive as to just how onerous transparent research processes are for both producers and consumers qualitative research.

Authors that were applying ATI to works in progress faced the additional challenge of deciding what constituted an annotation (in addition to the main narrative) and what content remained central to (and indeed a part of) the main narrative. In many instances authors of both completed and in-progress articles reported that they were conflicted as to how important annotations had become to interpreting their argument. Likewise, reviewers in both workshops commented on how important an annotation seemed that they questioned why such detail was left out of or not included in the main text. The complexity of deciding which narrative explanations of production or analytic transparency remain central to an argument, and what contextual information can be supplied to support an author's claims through annotation remains a challenge for future ATI implementations.

Incentives

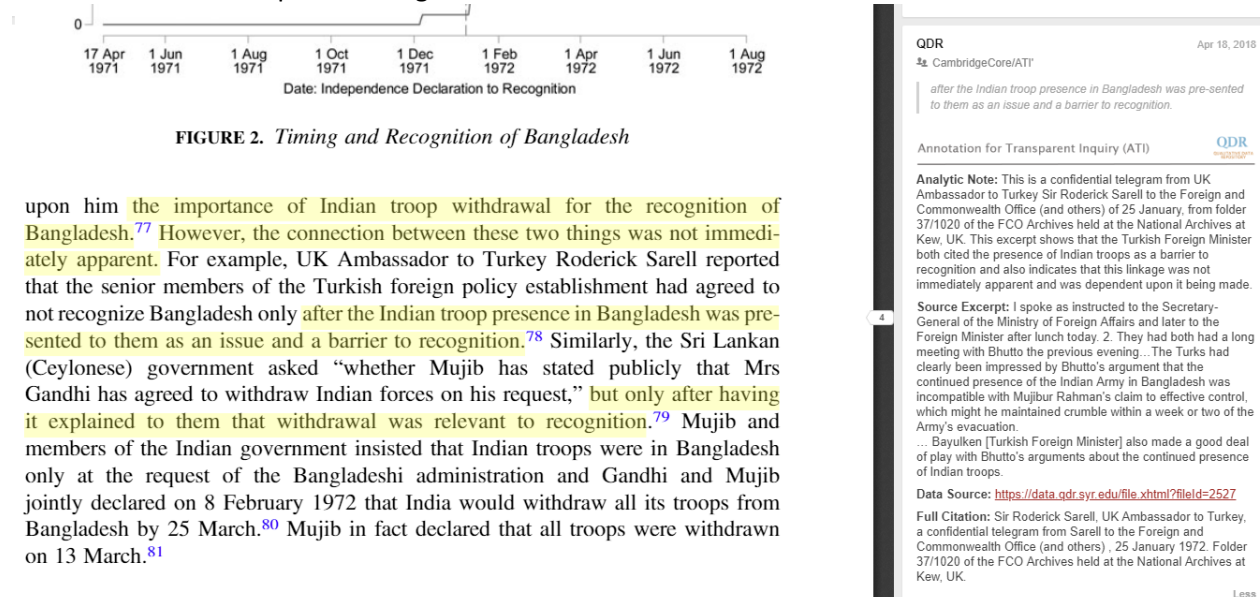
For producers of an ATI project, the labor needed to assemble data sources, develop concise and clear explanations of how data were collected, as well as produce analytic notes was similar in time and energy spent to produce the main article that was being annotated. Many authors, and their sympathetic reviewers, were concerned with the additional delay that producing an ATI project might introduce in already lengthy publication processes. For early career researchers, the potential delay of a publication was a major concern given the novelty of ATI. Authors participating in the second workshop faced the additional uncertainty of how annotations would be accessed and used as part of the peer-review process for getting their research published in a journal. Journal editors and reviewers will not likely be familiar with ATI, and this might add additional burden of educating editors and reviewers as to the importance and value of ATI as a supplement to a new manuscript submission.

Workshop participants did not question the overall value of producing and consuming ATI projects, but were unclear how these intense efforts would be rewarded. The challenge of rewarding open science practices is not unique to qualitative research (e.g. Nosek et al. 2015), but the incentive structure within this domain, like techniques for transparency more generally, is not yet well established. While incentives to reward transparent research efforts, like ATI, mature there are a number of practical short-term steps that can be taken to reduce the burden of creating ATI projects. We discuss these future steps in the technical challenges section below.

Signalling

Similar to the challenge of selecting which portions of an article to annotate, authors and reviewers also desired a typology of annotations that could act as a signal of the importance or function of an annotation to consumers. Workshop participants spent a considerable amount of time in focus groups discussing the merits, design, and potential application of an annotation typology. In the current implementation of ATI, a Hypothesis-supported annotation provides an in-line text highlight to indicate where an annotation should be anchored (see Figure 1 for an example). Authors and reviewers at each workshop desired additional signalling mechanisms, such as different colored highlights or tags attached to an annotation in order to alert a reader's attention to the importance and function of an annotated

passage. Reviewers in particular wanted a simple way to sort or sift through multiple annotations quickly to determine how important an author thought the annotation was to an argument, or more simply, to view all annotations that pertained to a particular mode of transparency (e.g. analytic transparency). There are existing vocabularies and data models that provide a scheme for this kind of contextual information about a web-annotation. Notably, the W3C Open Annotation Data Model provides a vocabulary for typing annotations by “motivation.” (Sanderson, Ciccurese, and Young 2017) ATI could implement or extend these “motivation” features so that authors may select from a controlled vocabulary to signal to readers what motivated their application of an annotation. This functionality however remains an open challenge for future ATI research.



Technical Challenges

The most significant technical challenge for ATI (and for annotations of scholarly content to date) is the ability to reliably display annotations. Hypothesis provides three mechanisms for displaying annotations:

1. Using an extension for the Chrome browser
2. Using a snippet of javascript code installed by the content owner (in our case: the publisher) on the pages to be annotated
3. By routing readers of annotations through a proxy server that reloads the requested page with annotations loaded and visible (for instance, via hypothes.is/iassistquarterly.com/ will load the IASSIST Quarterly homepage with the Hypothesis annotation bar on the left).

The first two options work reliably for scholarly content but are still comparably rare: Few users have the Hypothesis Chrome extension installed and the extension is currently neither available on any other browser nor on any mobile devices. Hypothesis annotations are currently enabled only by a small number of publishers (such as Ubiquity Press and eLife as well as OSF preprints) and was available for the majority of articles used during the first ATI workshop due to a collaboration with Cambridge University Press. Looking ahead, however, adoption by publishers appears to be increasing rapidly.

Given the limited availability of the first two options, the via.hypothes.is proxy currently plays a major role in making annotations visible. Unfortunately, however, it interferes with IP-based authentication,

which researchers use to access the vast majority of articles in the social sciences. Once routed through a proxy, a publisher no longer recognizes a request for an article as coming from University X, which has a subscription to the journal, and therefore does not display the full text of the requested article.

Next steps

There are a number of future research and technical developments that will ease the adoption and use of ATI by qualitative and mixed-methods researchers. As discussed above, the need to include annotations in a peer-review process may pose challenges in educating and familiarizing journal editors with the role of ATI in providing access to an author's supplementary data.

QDR staff will continue to closely examine the data collected during the two ATI workshops that encompasses logs and short surveys kept by authors, the feedback received from reviewers, as well as detailed notes from the two workshops that allow for a detailed initial evaluation of ATI. Using the results of this evaluation we expect to improve documentation and guide future technical innovations. A key area for future work that is emerging relates to bridging the divide between digital annotations and the fact that many researchers prefer reading physical paper (or paper-like) versions of publications.

Lastly, the labor of assembling an ATI project is undeniably a barrier to adoption. To overcome the challenge of transforming annotations that are produced in the typical writing workflow of an author (e.g. comments in a Word document), QDR has recently started to design a tool that will allow authors to upload their article (with comments) and data, then transform the comments into annotations that can be displayed using Hypothesis, and deposited with QDR for long-term preservation. This tool is in the early stages of development, but we believe it will be an important step in easing the process of assembling these diverse resources into a structured research object the creation and sharing of ATI projects.

References

- APA. (2016), "Ethical Principles of Psychologists and Code of Conduct", American Psychological Association, Washington D.C., available at <https://www.apa.org/ethics/code/ethics-code-2017.pdf>.
- APSA. (2012), *A Guide to Professional Ethics in Political Science*. 2nd ed., American Political Science Association, Washington, D.C., available at <http://www.apsanet.org/portals/54/Files/Publications/APSAEthicsGuide2012.pdf>.
- Couture, J. L., Blake, R.E., McDonald, G., and Ward, C.L. (2018), "A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing", Edited by Jelte M. Wicherts, *PLOS ONE*, vol. 13 no. 7, <https://doi.org/10.1371/journal.pone.0199789>.
- Crawford, T. (2015), "Data for: Pivotal Deterrence and the Chain Gang: Sir Edward Grey's Ambiguous Policy and the July Crisis, 1914", in *Pivotal deterrence: Third-party statecraft and the pursuit of peace*, QDR Main Collection, <https://doi.org/10.5064/F6G44N6S>.
- Freese, J. and King, M.M. (2018), "Institutionalizing Transparency", *Socius: Sociological Research for a Dynamic World*, vol. 4, <https://doi.org/10.1177/2378023117739216>.
- Geertz, C. (1973), *The Interpretation of Cultures: Selected Essays*, Basic Books, New York.

- Giofrè, D., Cumming, G., Fresc, L., Boedker, I., and Tressoldi, P. (2017), “The Influence of Journal Submission Guidelines on Authors’ Reporting of Statistics and Use of Open Research Practices.” *PLOS ONE*, vol. 12 no. 4, <https://doi.org/10.1371/journal.pone.0175583>.
- Hypothesis. (2019), *The Hypothesis Web-Based Annotation Client*. JavaScript, available at <https://github.com/hypothesis/client>.
- King, G. (1995), “Replication, Replication”, *PS: Political Science & Politics*, vol. 28 no. 3, 444–52.
- Kratz, J. and Strasser, C. (2014), “Data Publication Consensus and Controversies.” *F1000Research*, vol. 3 no. 94, <https://doi.org/10.12688/f1000research.3979.3>.
- Moravcsik, A. (2010), “Active Citation: A Precondition for Replicable Qualitative Research”, *PS: Political Science & Politics*, vol. 43 no. 1, 29–35, <https://doi.org/10.1017/S1049096510990781>.
- . (2014), “Transparency: The Revolution in Qualitative Research”, *PS: Political Science & Politics*, vol. 47 no. 1, 48–53.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S. et al. (2015), “Promoting an Open Research Culture”, *Science*, vol. 348 no. 6242, 1422–25, <https://doi.org/10.1126/science.aab2374>.
- O’Mahoney, J. (2017), “Making the Real: Rhetorical Adduction and the Bangladesh Liberation War”, *International Organization*, vol. 71 no. 2, 317–48, <https://doi.org/10.1017/S0020818317000054>.
- Sanderson, R., Ciccarese, P., and Young, B. (2017), “Web Annotation Data Model”, *W3C Recommendation*, World Wide Web Consortium, <https://www.w3.org/TR/annotation-model/>.
- Saunders, E. N. (2015), “Data for: John F. Kennedy”, in *Leaders at war: How presidents shape military interventions*, QDR Main Collection, <https://doi.org/10.5064/F68G8HMM>.
- Smith, J. and Holmes-Elliott, S. (2018), “Data for: The Unstoppable Glottal: Tracking Rapid Change in an Iconic British Variable”, Qualitative Data Repository, <https://doi.org/10.5064/F6OPE7MJ>.
- Udell, J. (2017), “Federating Annotations Using Digital Object Identifiers (DOIs)”, *Hypothesis Blog* (blog), June 22, 2017, available at <https://web.hypothes.is/blog/does/>.
- Walker-Peddakotla, A. (2018), “New Collaboration Capabilities for Annotation: Open and Restricted Groups”, *Hypothesis Blog* (blog), October 17, 2018, <https://web.hypothes.is/blog/expanding-our-groups-capabilities/>.

Endnotes

¹ Sebastian Karcher, Qualitative Data Repository, Syracuse University, skarcher@syr.edu

² Nicholas Weber, Information School, University of Washington, Seattle, nmweber@uw.edu; Authors listed in alphabetical order. The authors would like to thank their colleagues at the Qualitative Data Repositories as well as the authors and reviewers of the two ATI Initiative workshops.

³ Currently the conversion is performed manually, but automated conversion is technically feasible and planned.

⁴ Both of these features depend on the presence and accuracy of metadata in the site header, the citation_doi tag for the former and the citation_pdf_url tag for the latter functionality.

⁵ In many ways, this was a bias pool of researchers as they had all volunteered to participate based on their belief that not only was ATI valuable to their own research activities, but that helping to improve it through the workshop focus groups, might benefit later adopters.