# Linked Ethnographic Data: From Theory to Practice

Dominic DiFranzo
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY
difrad@rpi.edu

Marie Joan Kristine Gloria
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY
glorim@rpi.edu

James Hendler
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY
handler@cs.rpi.edu

## ABSTRACT
As Web Science continues to mix methods from the many disciplines that study the web, we must begin to seriously look at mixing and linking data across the Qualitative and Quantitative divide. A large difficulty in this is in modeling and archiving Qualitative data. In this paper, we outline what these difficulties are in detail with a focus on the data practices of Ethnography. We describe how linked data technologies can address these issues. We demonstrate this with a case study in modeling data from audio interviews that were taken in an ethnographic study conducted in our lab. We conclude with a discussion on future work that needs to be done to better equip researchers with these tools and methods.

## Categories and Subject Descriptors
H.1.1 Systems and Information Theory (E.4)

## General Terms
Human Factors; Theory

## Keywords
Ethnography; Linked Data; Web Science; Mixed Methods

## 1. INTRODUCTION
In trying to understand the Web (e.g. how it's used, constructed, etc.), current research often emphasizes the technological elements at the cost of the social, leaving the circle of Web Science research incomplete. This article moves to close the loop by examining emerging social concerns to inform the technical. Specifically, this paper builds on discussions for epistemological and methodological pluralism in systems used by researchers, like ethnographers and Web Scientists, who use the Web as both an artifact *of* study and *for* study.

In this paper we explore why leveraging linked data principles can enhance a qualitative researcher's workflow through provenance and metadata. We present our application as a case study focused on protest rally participant behaviors. As the focus of this paper is to discuss alternative tools, the findings from the case study are omitted as they fall outside the scope of this paper. We conclude by reflecting on the value of Linked Data technologies in qualitative research, and the need for further theoretical and technical development in this space.


*WWW'15 Companion*, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745942

## 2. Case Study
The following section demonstrates our application of linked data to a specific use-case ethnography study being conducted in the Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute. To reiterate, the goal of this section is to highlight the method and process of which linked data is leveraged to encode data captured through more traditional ethnographic techniques. As such, final analysis and commentary regarding the privacy study will not be included as it falls outside the scope of this piece. The goal, however, is to showcase the potential impact linked data technologies may have for more traditional qualitative research methods.

Early exploratory fieldwork was conducted in fall 2013 as part of a longitudinal study investigating individual constructions and practices of privacy. The field site was located in Washington, D.C. and motivated by an anti-surveillance protest rally. Prior to the event, the researcher engaged in the development of semi-structured questions based on relevant domain literature to elicit answers regarding personal motivation. Sample questions include (but are not limited to): Why did you attend today's event?; How did you hear about today's event?; and How knowledgeable are you about the concerns related to the event? The nature of these questions were intentionally open-ended in order to capture as much information as possible from each interview. Subjects were chosen at random, on site during the rally (though a self-selection bias is noted as it was an opt-in participation). All interviewees were explicitly informed of the nature of which the information would be used in the future and were given the opportunity to provide additional identifiable information, like personal emails, if desired.

All interviews were captured via audio recording (the "voice memo" iPhone app), which were later transcribed and coded. The audio files have been transferred and are stored on a local machine. No personal identifiable information was audible within any of the recordings and all participants remain anonymous throughout the capture, transcription and coding of the audio data. In addition to the audio interviews, pictures and video of the entire rally were also captured by an associated researcher. However, this data has yet to be annotated or coded. The coding parameters reflect ideas conveyed in the literature which spans multiple domains from surveillance studies, cognitive science, psychology, legal studies and media studies.

Taking the table of the transcription of the interview, along with the coding, we were able to convert this information to linked data using the converter (CSV2RDF4LOD)[1] developed by the TWC for the Open Government Data project. This converter

[1] https://github.com/timrdf/csv2rdf4lod-automation/wiki

can take any tabular data and convert this to RDF, and remodel it using any vocabularies or ontologies. the Open Annotation Data Model (http://www.openannotation.org/spec/core/) to model the annotation (in our case the transcription is the annotation). Additionally, we included coding of these transcriptions from the interview. In Figure 1, an example outline of one such transcription annotation is provided. This annotation is linked to the metadata that describes when it took place, who wrote the annotation, etc. With this structure, other researchers can provide their own annotation and transcription to this same audio section, link it to the specific audio URI and preserve his or her own interpretation or understanding of the same artifact. Additionally, these annotations can be linked to other data and media assets. For example, this annotation in Figure 1 could be linked to a tweet that took place at the same event and was coded in the same manner by the original or other researchers.

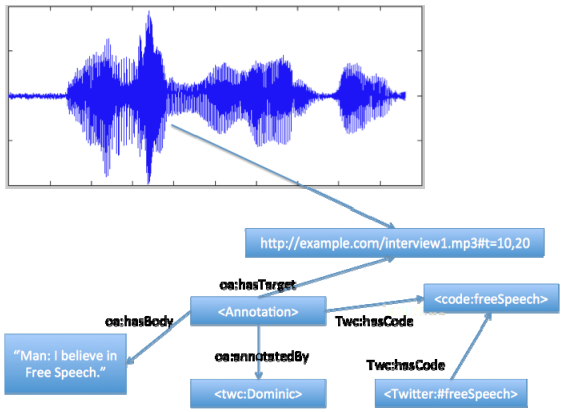


**Figure 1: Model of an audio segment being transcribed and coded by a research, using RDF**

## 2.1 DISCUSSION

While this annotation tool is an interesting first step in utilizing linked data and semantic web technologies for structuring and modeling ethnographic data, much more work still remains. First and foremost is the need for better tools and user interfaces that allow non-technical qualitative researchers to produce linked data on their own. Our example of the RDF encoding was only successful because of the research team's technical proficiency and familiarity with the tools used. Additionally, one must have working knowledge of data modeling and semantic web technologies. The key is to develop and integrate these semantic technologies into tools and frameworks already used by social scientists.

We also recognize that in enabling such collaborations, safeguards to protect the privacy of the data will be critical. As most human-centric research requires institutional approval, considerations for how certain data parameters and consent should be deliberated. For example, we ask questions of *how much and what can be shared?; how is anonymity maintained?; how do researchers evaluate risks for subjects if data is shared?;* and, *how can we leverage linked data to protect and control this information?*.

# 3. REFERENCES


1. Angrosino, Michael V., and K. A. Mays de Pérez. Rethinking observation. From method to context. Teoksessa NK Denzin & YS Lincoln (toim.) Handbook of qualitative research. Thousand Oaks: Sage, 2000.
2. Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences. MIT press.
3. Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679.
4. Dublin Core Metadata Element Set, Version 1.1: Reference Description, Dublin Core
5. Metadata Initiative, Reference Description 2004-12-20 2004.
6. DCMI Usage Board. The Dublin Core Metadata Initiative. http://dublincore.org/documents/dcmi-terms/
7. Gloria, M. J. K., DiFranzo, D., Navarro, M. F., & Hendler, J. (2013, May). The performativity of data: reconceptualizing the web of data. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 109-117). ACM.
8. Halford, S., Pope, C., & Weal, M. (2013). Digital futures? Sociological challenges and opportunities in the emergent Semantic Web. Sociology, 47(1), 173-189.
9. Hooper, Clare J., Nicolas Marie, and Evangelos Kalampokis. "Dissecting the butterfly: representation of disciplines publishing at the web science conference series." Proceedings of the 3rd Annual ACM Web Science Conference. ACM, 2012.
10. Kramer, S., et al. "Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model." (2012).
11. Poirier, L., DiFranzo, D., Gloria, M. J. K., (2014, June). Light structure in the Platform for Experimental Collaborative Ethnography. In Web Science 2014 Workshop Interdisciplinary Coups to Calamities
12. Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., & Hendler, J. (2013). The Web Science Observatory. IEEE Intelligent Systems, 28(2), 100-104.