Opaque Data Citation: Actual Citation Practice and Its Implication for Tracking Data Use

Elizabeth Moss (eammoss@umich.edu) & Jared Lyle (lyle@umich.edu), ICPSR, University of Michigan

What is the ICPSR Bibliography of Data-related Literature?

ICPSR is a social and behavioral science data repository in the United States, which curates and disseminates over 10,000 data collections for re-use. ICPSR also tracks published analyses of those data in the ICPSR Bibliography of Data-related Literature. A searchable database, it provides two-way links between data archived at ICPSR and over 75,000 scholarly publications.

The biggest obstacle to tracking data use

In the social science literature, most data attribution is incomplete and does not include persistent identifiers (PIDs). Instead, authors mention data opaquely. Without explicit data citation, a publication cannot automatically or definitively link to a data source. The human effort required to find, interpret, and link opaque citations is costly and inefficient, so data use often goes untracked, and data creators go uncredited.

Opaque data citation may be:

Almost complete. The data are formally cited in the references, but the PID is not included, which would have enabled machine-actionable detection and linking.

> US Dept. of Justice, Federal Bureau of Investigation. (2003). Uniform Crime Reporting program data [United States]: Arrests by age, sex, and race, 2003 [Computer file]. ICPSR04285-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2007-03-21.

Indirect. A title and year of the data are mentioned in the methods, so it is clear data were used and even named, but no formal reference is made to the data or where they may be accessed. Often, the reference provided is to another publication and not directly to the data.

> "Our analysis draws upon the India Human Development Survey 2005 (IHDS) collected by the University of Maryland and the National Council of Applied Economic Research (NCAER) in 2005."[46]
>
> [46] Desai S, Dubey A, Joshi BL, Sen M, Shariff A, Vanneman R. Human Development in India: Challenges for a Society in Transition, New Delhi, India: Oxford University Press, 2010.

Mismatched. When the data are archived, the title often differs from the title used in publications prior to archiving.

You see it in the literature this way:	Title used when archived:
"Data analysed from WHO World Health Survey (WHS)."	WHO Study on Global AGEing and Adult Health (SAGE): Wave 1, 2007-2010 (ICPSR 31381)

Barely there. Authors may mention the name of a data collection, and may even provide the specific years analyzed. But they do not include the PID provided by the archive, let alone the version number.

> "This study investigated the association between victim reporting and the police response to past victimizations with data from the National Crime Victimization Survey from 1998–2000."

Vaguely described by necessity. The investigator writes about her own data well before having a citation and registered PID. Or, oftentimes, in sensitive areas of research, authors tend to say as little as possible about identifying characteristics of the data, even in broad terms.

> "Data were collected from 1,342 men in First Offender programs in California (n = 996), Oregon (n = 77), and Nevada (n = 269). Men voluntarily participated in these deferred adjudication programs following their arrests for attempting to hire a prostituted woman on the street."

Deducible with inside information. In some cases, even though no formal citation is used, an informed reader may know where the data were mandated to be deposited. In other cases, authors acknowledge assistance from a repository.

> "These findings emerge from a two-year study that focused on Muslim-Americans in four communities: Seattle, Washington; Houston, Texas; Buffalo, New York; and Raleigh/Durham, North Carolina."

Dealing with the opaque reality

Rethink how we educate about the importance of good citation practice, especially with an understanding of opaque citation in mind. *Why do authors continue to opaquely cite data? Should we rethink the format and method of guidance and enforcing compliance? Are there new types of training that would nudge an author to formally cite data?*

Machine learning and natural language processing hold much promise in improving data usage tracking for opaque citations. However, we first need to understand current practice in order to create and train algorithms. Training sets could be developed with the help of curated collections like the Bibliography.