Realistically, Isn't Everything Social Science Data?

Cindy Severt, Senior Special Librarian, University of Wisconsin, Madison

International Federation of Library Associations, Social Science Libraries Section, Satellite Conference

Disappearing disciplinary borders in the social science library - global studies or sea change?

University of Toronto, Toronto, Canada
6-7 August 2008

Sponsored by:

CQ Press
Worldwide Universities Network (WUN)
Center for Global Studies (CGS) at the University of Illinois
University of Illinois Library
University of Toronto Libraries
Faculty Information Studies, University of Toronto

Introduction

As an academic data librarian serving primarily traditional quantitative social science departments since 1990, I started noticing about seven years ago that not only were the types of requests I was getting changing; my user group was taking on a decidedly different profile. It was after an economics student remarked, "Isn't everything social science data if it can be analyzed as such?" that it occurred to me to wonder, "Exactly what *is* social science data? And what is it not?"

Traditional vs. New Users and New Research Interests

Because large numeric quantitative datasets must be analyzed with statistical software, its practitioners and users have tended – until recently -- to be those with a background in statistics: demographers, sociologists, economists, and a sprinkling of political scientists. Add undergraduates charged with having to incorporate numeric data for a research paper into the mix, and suddenly you get a new generation of users with generationally different research interests. Once a new user group appears on the horizon, others follow from such unlikely departments as Humanities, History, Business Administration, Medicine/Dentistry, Jewish Studies, Psychology, Law/Legal Services, and Nursing, especially if they can come to your door without leaving their office. The School of Journalism & Mass Communication is on the verge of rivaling the department of sociology as my largest user group. But it wasn't until a team of researchers from the Center for Sustainability and the Global Environment showed up at my literal door seeking spatially explicit datasets on the global distribution of crops and their correlation with the spatial distribution of observed fires from satellites that I sensed I was in unfamiliar territory. That sensation was reinforced a few weeks later when I received a request for publicly available data on the number of human trafficking victims world-wide by country from an International Public Affairs student.

Traditional vs. New Forms of Data

In certain respects microdata is the perfect tool for interdisciplinary research since conducting

secondary analysis is to analyze data for a purpose other than that for which it was originally

collected. But how reliable is questionnaire-derived data in this day and age when we are

overwhelmed with surveys that seem to ask the same questions over and over again, short on

time, fear our privacy is being raided and our identity is being stolen, and either incensed or

blasé about the fact that *"they"* already know so much about us? It's only as reliable as the

respondent cares to be. It's easy for respondents to gloss over or ignore questions simply to

get the task over with; or conversely, to exaggerate, fabricate, or intentionally skew results

either out of annoyance, or simply because they can. To quote a colleague who was a recipient

of the American Community Survey:

> *It was about 17-18 pages, if memory serves. There were over fifty*
>
> *questions. I didn't really fabricate responses I simply estimated them*
>
> *particularly with respect to questions about income. So while I*
>
> *suppose that the responses I gave were "in the ballpark" they were*
>
> *in no way exact. My question is how many other respondents use*
>
> *the same "technique" that I do, and how does this affect the accuracy*
>
> *of the survey?*

An expanded user group has resulted not only in a wider range of reference questions being

asked, but in new forms of data being analyzed to answer those questions, as well as new (and

sometimes unexpected) formats for delivering such data.

++

Social Networking Sites

A December 17, 2007 article in the *New York Times* posits that it's Facebook's role as a Petri

dish for the social sciences — sociology, psychology and political science — that particularly

excites some scholars, because the site lets them examine how people, especially young

people, are connected to one another, something few data sets offer……"One of the holy grails

of social science is the degree to which taste determines friendship, or to which friendship

determines taste,"[1] said Jason Kaufman, an associate professor of sociology at Harvard

University and a member of a Harvard/UCLA research team studying the question: Do birds of a

feather flock together, or do you become more like your friends? In a blow to internet

utopianists, preliminary findings indicate that birds of a feather *do* flock together. While race

and gender have the largest influence on who one befriends in social networks online, we tend

to form friendships with others who are like us. The project, ongoing through 2009 aims to build

a longitudinal data set comprising a complete cohort of college students' social network and

cultural preference data from their freshman through senior years.

Biosocial Data

Biosocial data are population based sample surveys that combine demographic, social, and

behavioral data with biological indicators. Most biosocial data collects not DNA, but biosocial

markers long associated with health surveys such as grip strength, pulmonary functioning,

clinical measurements of various substances in the blood, saliva, or urine, blood pressure, heart

rate variability, weight and height, perceived age, and various other physiological and

---

[1]Rosenbloom, Stephanie. "On Facebook, Scholars Link Up With Data". *The New York Times.*
December 17, 2002.

anthropometric measures of risk factors, exposures, and health outcomes which social scientists can use to better estimate environmental and behavioral effects on health.[2] When DNA *is* collected genetic variations become variables, inevitably leading to questions such as whether or not genetic indicators should be analyzed in the context of understanding complex social traits. At what point do sociology and behavioral/natural science overlap? Are human behavior and culture a product of natural selection?

Audio/Video Data

What if the variables in social science datasets weren't limited to the survey instrument? What if researchers could create their own variables from raw video footage? TRANSANA software for analyzing digital video or audio data does just that. Developed by a graduate student and maintained at the Wisconsin Center for Education Research, TRANSANA works by coding events in video clips and linking those events to an audio transcript. To illustrate, a researcher studying how children learn in a classroom setting might notice recurring episodes of laughter during a 60 minute video. Episodes of laughter can be tagged to a transcript and selected, not unlike selecting variables by column location. Episodes of laugher, silence, activity, even facial expression can be "extracted" and exported as a tab-delimited file for analyzing with statistical software. What happens just prior to and after the laughter might be equally important and it is here that TRANSANA can leverage the richness of context. (www.transana.org)

Virtual Worlds

---

[2]Maxine Weinstein, James W. Vaupel, and Kenneth W. Wachter. Eds. *Biosocial Surveys,* p.3. The National Academies Press: Washington DC, 2008

Societies are not (so) amenable to experimentation. Government economists would not be popular if they repeatedly tweaked monetary policy just to see what happened. Social scientists learn from history. They run surveys. They even conduct small experiments with a handful of subjects in idealized conditions. What they can't do is manipulate the system they are studying.[3]

Or can they? Edward Castronova, an economist at Indiana University in Bloomington, and a team at the Synthetic Worlds Institute intend to find out. Named after the forest of William Shakespeare's *As You Like It* in which love in its many guises is explored, ARDEN http://swi.indiana.edu/ardenworld.htm is a multi-player online role-playing game that attempts to immerse users in the narrative, language, and culture of Elizabethan England; sort of an *EverQuest* meets *MacBeth*.   Its objective is to use virtual worlds as laboratories for research on macro-level social phenomena. Rather than simulating real life, virtual societies are created with the intention of deliberating tweaking them by tinkering with monetary policy, rewriting the rules of democracy, forcing players to work together or driving them apart. Each social experiment is conducted systematically holding all other variables constant and rigorously monitoring the outcomes. An artificial *Truman Show* if that's not too redundant.

Jeremy Bailenson, head of the VIRTUAL HUMAN INTERACTION LAB, http://vhil.stanford.edu/ and an assistant professor of communication at Stanford University studies the way self-perception affects behavior. No surprise that what we think about ourselves affects the confidence with which we approach the world. What is a surprise is that this applies in the virtual world too….. Bailenson's research (also) suggests that the qualities you acquire online — whether it's confidence or insecurity — can spill over and change your conduct in the real world,

---

[3] Castronova, Edward. "Arden Experimental Results: Preliminary Results". 16 May 2008. Terra Nova blog.
http://terranova.blogs.com/terra_nova/2008/05/arden-experimen.html

often without your awareness…………The possibilities are virtually endless: Inhabit buffed-up

versions of yourself to lose weight, cuter versions of yourself to gain confidence, or older

versions to start putting money away for the future.[4]

Using immersive virtual reality, it is possible for someone to literally experience the world as

another person; to walk a mile in the shoes of another. Bailenson's team is using these

simulations to explore relationships of gender, status, and race, and testing the theory that

wearing the face of another in a simulation designed to highlight diversity issues can increase

awareness.

Data Visualization Tools (Open Source using Web 2.0 functionality)

Data Visualization is the set of techniques used to turn a set of data into visual insight. It aims to

give the data a meaningful representation by exploiting the powerful discerning capabilities of

the human eye*.*[5]   In other words, it's transposing numbers into pictures; repackaging existing

data from something statistically extracted to something visually comprehensible.

In his thorough paper *DISC-UK DataShare: Web 2.0 Data Visualization Tools: Part 1 –*

Numeric Data[6] Stuart MacDonald writes, As Web 2.0 continues to evolve and transform

into Web 3.0, we are seeing the boundaries between websites and web services

---

[4] Dell, Kristina. "How Second Life Affects Real Life". *TIME.* 12 May 2008.
http://www.time.com/time/health/article/0,8599,1739601,00.html

[5]*Edinburgh Online Graphics Dictionary*. Ed. Robert Fisher. 12 September 1999.
http://homepages.inf.ed.ac.uk/rbf/GRDICT/grdict.htm

[6]*DISC-UK DataShare: Web 2.0 Data Visualization Tools: Part 1 – Numeric Data*. Stuart MacDonald.
14 January 2008. Joint Information Systems Committee Repositories and Preservation Programme:
Repositories Start-up and Enhancement Projects Strand. http://www.disc-uk.org/docs/Numeric_data_mashup.pdf

blurring, as more and more web content becomes remixable. A quick overview of some

of the some data visualization tools covered in his paper includes:

MANY EYES

http://services.alphaworks.ibm.com/manyeyes/home

This IBM utility aims to "democratize" visualization and to enable a new social kind of

data analysis. Many Eyes allows registered users to upload their own data on the

understanding that the data is made available to all. It also has the facility to visualize

numeric data in a spatial manner (using the proprietary mapping utility World Map).

Users can view (or create) topic hubs which are pages devoted to a particular topic or

theme. It serves as a way to collect related visualizations and data sets.

SWIVEL

http://www.swivel.com/

Swivel aims to 'liberate the world's data and make it useful so new insights can be discovered

and shared'. Swivel charges a fee for any restrictions a user may impose on the data. Swivel

Geography or Swivel G allows users to zoom in and out and move seamlessly from one part of

the globe to another. Swivel G also includes interactive features that enable users to drill down

into specific data points in more detail.

GRAPHWISE

http://www.graphwise.com/

Graphwise describes itself as 'a beta release of a search engine for table data'. Upon

registration, users can search for and upload tabular data from their local machine or 'scraped'

from an internet web page in a number of formats including Excel (.xls), comma delimited (.csv),

HTML (.html), PDF (.pdf), Lotus spreadsheets.

To Serve, or Serve Not

Just as the lines between traditional disciplinary boundaries have been blurred, so have the

lines between quantitative and qualitative data been blurred as diverse types of data are being

subject to analysis. In this increasingly complex and technological world, sociology continues to

cross pollinate and subdivide.

Spearheading the effort to serve the data needs of interdisciplinary researchers should be a

service and funding model designed for an entire campus not just a special interest group; and

a data services division that is visible to the entire campus community and not buried in a social

science department. Other anticipated outcomes:

- Traditional data librarians will see more and more undergraduates and will need to
  broaden their definition of what "data" is.

- More departments will have need for statistical consulting.

- Traditional reference departments will become more data-savvy as the line between
  what is, and what is not data, blurs.

- The once deep and narrow data well will be become increasingly shallow and wide.

- Need for more global (less regional) datasets such as the Global Terrorism
  Database, the International Social Survey Program, the World Values Survey,
  the United Nations World Crime Surveys.

Appendix

International Federation of Library Associations, Social Science Libraries Section, Satellite Conference, *Disappearing disciplinary borders in the social science library - global studies or sea change?* University of Toronto, Toronto, Canada, 6-7, August 2008

The Arden Project recently conducted a month long experiment, the results of which will be released during the International Communications Association meetings in Montreal the weekend of May 23rd. The experiment addressed the question: Are fantasy game players economically "normal"? Or on the contrary, when they make themselves into elves and dwarves and hobbits, do they stop taking economic decisions seriously? On the Terra Nova blog http://terranova.blogs.com/terra_nova/2008/05/arden-experimen.html#more Mr. Castronova wrote, "We created two virtual worlds, one an exact copy of the other, except that in the experimental world the price of a simple healing potion was twice as high as in the control. If people are taking prices seriously in this fantasy environment, they should buy fewer of the potions when potions are more expensive." Results were unavailable as of the submission of this paper.