Tools, Toys, and Filters

A Tinker's Apology

Bryan Jurish

"Use filters" —Brian Eno & Peter Schmidt, *Oblique Strategies*, 1975

As a tinker of algorithms, a tweaker of data structures, and a dyed-in-the-wool Platonist, I am committed to the (objective) existence of mathematical entities such as numbers and the relations between them. I nonetheless follow Christiane Birr in her skepticism regarding Anderson's (2008) blithe assertion that "given enough data, the numbers speak for themselves". Numbers seldom lie *per se*, but neither are they renowned for their loquacity. The traditional distinction between deductive truths about formal objects and inductive interpretations of empirical data acquired from a data sample such as a text corpus is useful here. Computers are well-suited to deductive tasks involving counting and other numerical manipulations, and are quite reliable for such purposes. They are not very adept at e.g. deciding *what* to count (corpus selection) or drawing (creative, interpretative) conclusions based on (numerical) data; as Silke Schwandt suggested: "I cannot find what I didn't look for", and "the interpretative act is my own". I submit here the proposition that computational tools for humanities research ("DH") are best understood as *filters* in the sense of Shannon's (1948) model of communication, also cited in the current context by Manfred Thaller.

In terms of Shannon's model, we should first acknowledge that natural language itself is a "lossy" or "noisy" encoding/decoding scheme ("codec"): ambiguity, underspecification, and other opportunities for misinterpretation abound in linguistic communication (Reddy, 1979). DH tools acting on text data typically compress the (already error-laden) signal further by applying a tool-specific data model (e.g. word counts), performing formal manipulations on that representation, and formatting the results for human inspection. In terms of Shannon's model, this is simply an additional encoding applied to the (already text-encoded) original message, i.e. a filter. A "lossy" filter degrades messages passed through it: most exploratory DH tools fall into this category, since implicit in their design is a desire for high compression rates on the one hand (we already have the text-encoding), and on the other because a precise characterization of the formal models required for a 1:1 reproduction of the original (semantic, communicative-intentional, transmitter-internal) message has thus far eluded us (and possibly always will).

Lossy filters should not disturb us, however: as humans, we come equipped with (are predisposed to) a whole bevy of integrated filters: linguistic filters for parsing (minimal attachment) and interpretation (semantic priming), perceptual ones for motion detection and voice recognition, cognitive filters for object independence and causal relations, as well as cultural ones for shared experience and common knowledge. Adding another (lossy) filter to our data intake process increases the informational "distance" in Moretti's sense, but does not change the fact that the

communication channel between the transmitter (text, author, object) and the receiver (ourselves, subjects, minds) is already noisy (i.e. fallible). The "intuitivity" often predicated of DH tools is nothing more or less than an exploitation of the human users' pre-existing perceptual/cognitive/cultural filters by use of color, motion, size, or shared metaphors such as tag-clouds, time series or histogram plots, etc. Such exploitation can be considered successful to the extent that all and only the *relevant* data is passed through both the programmatic and user-integrated filters, e.g. when the most "interesting" feature of the data is the most visually striking element of the presentation format.

The validity of interpretative conclusions drawn from empirical input is a well-known epistemological problem (induction); the question for DH is whether or not we are willing to accept yet another layer of filters on the data we consume. There are good reasons to do so. Perceptual filters tend to act as a "fast lane" for salient environmental data: visual sensitivity to motion for example can alert us to the potential presence of a predator. We retain the option of subsequently redirecting our conscious attention to the detected phenomenon for detailed inspection and interpretation ("was the motion caused by a hungry tiger or a frightened rabbit?"). Exploratory DH tools can act similarly as a "fast lane" for salient cultural data, constructed to facilitate subsequent refocusing on a detailed inspection (close reading) of "interesting" phenomena – where "interest" is a function of the user's individual research program. DH tools need not replace traditional close readings, but can instead act as "coarse caricatures" or "executive summaries" indicating which (textual) phenomena might warrant more careful study. As tinkers, our task is to minimize the *apprehended* lossiness of the filters by optimizing our data models & manipulations for the users' common research goals, analogous to the optimization of popular audio codecs (e.g. mp3, ogg) for the human auditory perceptual apparatus. This can be a frustrating task, since the research goals of humanities scholars can vary widely, and commonalities can be difficult to identify and formally characterize. As noted by various colleagues, communication and compromise between humanities scholars and tool builders working together is the most promising path for improvement in this regard.

Implicit above is the assumption that use of computational tools does not itself affect humanities scholars' underlying research goals. I propose that DH methods *can* however alter the tempo and spirit of (certain aspects of) the humanities research process: speedy responses and intuitive (exploitative) interfaces can allow a "playful" interaction with the underlying data and rapid ("agile") adaptation of (potential, proto-) research questions in response to the (real, "objective") formal properties of the sample as encoded by the method in question. Here again, the key element is the cohesion of the tool codec (data model and presentation format), the user's research interests, and his or her pre-existing perceptual/cognitive filters: playful interaction implies that I as a user am open to distraction and continuous creative re-invention of the activity at hand, which means I must have sufficient cognitive resources available for re-allocation. If I can rely on my integrated perceptual/cognitive apparatus to inform me of "interesting" phenomena – if the programmatic, scholarly, and integrated perceptual/cognitive filters cohere – then I likely have such resources available; otherwise, distractions tend to be simply "irritating".

As a final observation, the issue of cohesion is also of central importance to my own work as a

builder of computational tools. These must be evaluated on at least two independent scales: intrinsic properties such as correctness and complexity can be formally evaluated and discussed in the methodological domain (computer science, computational linguistics, etc.). As *tools*, they must also be evaluated in terms of extrinsic properties such as flexibility and utility, which are only predicable relative to one or more given user-dependent tasks. I as a tinker therefore humbly ask for the help, patience, and cooperation of curious humanities scholars, that together we might develop less irritating, less restrictive, more interesting, and more coherent tools (and toys).

References

Anderson, C. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired*, 23 June, 2008.

Reddy, M. J. "The conduit metaphor: A case of frame conflict in our language about language". In A. Ortony (ed.), *Metaphor and Thought*, pp. 284–310. Cambridge University Press, 1979.

Shannon, C. E. "A mathematical theory of communication". Bell System Technical Journal, 27(3):379–423, 1948.